4-19-2017

Demographic Characteristics and Measures of Teacher Performance in Urban Schools

Tequilla Andrea Banks

DEMOGRAPHIC CHARACTERISTICS AND MEASURES OF TEACHER PERFORMANCE IN URBAN SCHOOLS

by

Tequilla Andrea Banks

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Education

Major: Leadership and Policy Studies

The University of Memphis

May 2017

Abstract

Policy shifts over the last decade have resulted in an increased focus on teacher effectiveness as a key lever for increasing student academic outcomes. As a result, districts and states began overhauling their teacher evaluation systems to more accurately assess the performance of teachers. Many of these models included multiple measures that when combined, are believed to more accurately measure a teacher's individual effectiveness. Because these models are being used to make human capital decisions, it is imperative that the models be examined for both their efficacy and lack of bias.

Ultimately, this study examined two overarching themes: whether the teacher evaluation model utilized in one large urban district provides an accurate assessment of teacher quality and whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study was conducted to determine whether the teacher evaluation system being examined, accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve.

The data revealed that relationships existed between the three primary components of the teacher evaluation system being examined: classroom observations, student perceptions, and value-added, or growth scores, indicating that it was in fact an accurate method for assessing teacher performance. However, the study revealed relationships between the components of the model and characteristics of teachers and schools. School culture and poverty concentration were linked to teacher performance ratings. Results also showed that certain demographics, such as teacher race and school

poverty concentration were in fact predictive of effectiveness ratings. The study found that some teacher and school characteristics did predict teacher performance.

Implications resulting from the study should lead district leaders to consider how evaluation scores are interpreted for certain races of teachers, particularly when these teachers are serving in more challenging school environments (across-school variance) and serving at-risk populations of students (within-school variance). Additional analyses should be conducted to further investigate the unmitigated effects of these variables in influencing a teacher's performance.

Table of Contents

Table of Contents

List of Tables

Chapter 1: Introduction

The United States is in crisis. The achievement of U.S. students in math, literacy, and science lags behind the performance of students in many other nations. As evidenced by the Program for International Student Assessment (PISA) study in 2012, which examined the performance of 15-year-old students across these content areas, U.S students scored lower than the average for all the Organisation for Economic Co-operation and Development (OECD) countries in mathematics. While the scores were not significantly different than the average in reading and science, U.S. students still lagged behind students in 19 and 22 education systems, respectively (NCES, 2015). The PISA study is not the only metric used to assess how U.S. students are performing. NAEP, which is our nation's yardstick for measuring how well our students are performing, reveals similar statistics.

Even after transitioning to the more rigorous Common Core State Standards (CCSS), student proficiency in math declined while reading performance remained the same. More specifically, only 35% of our nation's fourth and eighth graders tested in 2015 were proficient in reading while 37% were proficient in math. Additionally, SAT scores dropped significantly; a decline of 7 points in one year (Adams, 2015). These data clearly indicate that many of our students are not well-prepared for success beyond high school, in college and career. Despite state assessments showing that students are performing well, nearly 60% of students must enroll in remedial coursework upon enrolling in postsecondary institutions (SREB, 2010). It is imperative that we identify factors contributing to the substandard performance of our nation's students.

One such factor that has been shown to explain a significant amount of variance in student outcomes is the performance of the teacher. In fact, research has shown that the teacher is the single most important determinant of student achievement (Rivkin, Hanushek, & Kain, 2005). Even when students hail from economically disadvantaged households, an effective teacher can facilitate them in making gains comparable to those peers from higher socioeconomic backgrounds. Contrastingly, a teacher who is marginally effective or ineffective cannot only negatively impact a student's trajectory in K-12 but also impact their lifetime earning potential. A research study conducted by Chetty, Friedman, & Rockoff (2011), where students were essentially tracked from elementary school to their careers, revealed that students consistently taught by teachers who facilitate student growth over the years are more likely to attend college and less likely to become pregnant than their peers. To put this more concretely, a classroom of students taught by a highly effective teacher, or a teacher in the top 5% who constantly pushes the needle on academic growth, will earn about $250,000 more during their lifetime than a class of students who were not afforded the same opportunity (Chetty et al., 2011).

While all students need access to effective teachers, the impact of ineffective teaching practices on students from disadvantaged backgrounds is even more concerning. Students from lower socioeconomic backgrounds often achieve at lower academic levels than their peers from higher socioeconomic backgrounds. While our national graduation rate is at its highest, students from the most disadvantaged backgrounds are not experiencing the same success as their peers. In fact, high poverty students graduate at

lower rates than their peers in every state for which statistics were available. Of our 50 states, low income students only performed better than their less disadvantaged peers in six states (Cosman, 2014). NAEP results revealed similar disparities. In 2015, only 21% of fourth grade students from low-income backgrounds achieved proficiency on the reading test compared to more than half of their peers (Boser, Baffour, & Vela, 2016).

While some have attributed this difference in performance to the households in which many students from lower socioeconomic backgrounds are raised, others claim these differences can only be partially attributed to the lack of resources in disadvantaged homes, such as computers and high-quality reading materials (Lubienski & Lubienski, 2005). While it is obvious that schools cannot control a parent's ability to provide their children with material resources, schools can control the teachers who are hired and retained to educate students every day. The disparity in educational outcomes for students from disadvantaged backgrounds necessitates efforts that place the most effective teachers in schools where they are most needed.

While providing the most effective teachers to the students who need them most is important, many districts and states have found that it isn't the easiest task to accomplish. In fact, districts have attempted to implement practices to attract high-quality teachers to more disadvantaged schools. These practices include additional monetary incentives and even allowing teachers to transfer with colleagues from their schools to new schools. In most cases, teachers just aren't willing to do so. Working in schools with high-poverty students is more challenging and requires a unique skillset that many teachers don't yet have. There are some high-quality teachers who have left their schools

in less impoverished areas to teach in high poverty schools, however. As revealed by Papay (2013), these teachers often become disillusioned. While teachers prefer to work in schools where they have access to a variety of resources that they can use to facilitate success for students, teachers are not exiting due to a lack of resources. Instead, factors such as school culture, peer relationships, and support from school leadership often drive teachers to leave these schools. Failing to build strong cultures and leadership in high poverty schools results in a revolving door, where teachers with potential constantly exit and are replaced by teachers who are less effective and more inexperienced.

According to Taylor (2005), additional inequities in educational institutions are attributing to the poor performance of students from disadvantaged backgrounds. Teachers in high-poverty schools don't receive the training and support that they need to improve. Due to a lack of school funding, these teachers aren't able to engage in professional development activities as much as they need to. School leaders are often so focused on student discipline and management that they are unable to fully commit to serving as instructional leaders in their buildings. Since these teachers are already less experienced in most cases, their growth trajectories are often stunted as a result. This widely seen set of practices results in less qualified teachers being more prominent in schools that represent disadvantaged student populations. When the neediest students are not provided with access to effective teachers, the result is substandard performance (Burris & Heubert, 2006).

Background of the Problem

For decades, the methods used to identify high quality teachers, the teacher evaluation process, were ineffective. Although many students were experiencing academic difficulty, counterintuitively, most teachers were consistently rated as effective or highly effective by their school principals. In fact, The Widget Effect, a study (2009) conducted by The New Teacher Project, TNTP, in 12 school districts, revealed that less than 1% of teachers received unsatisfactory ratings. In many cases, school districts assessments of teacher performance were based on only 60 min of instruction a year although teachers are with their students for 180 days. In some states, teachers were observed even less regularly. In Tennessee, after teachers received tenure, principals were only required to observe them once every five years.

Even when teachers were observed regularly, most districts did not use data to improve instructional practice. For example, principals in the district being studied were not required to have conversations with teachers about their observation results or to provide feedback based on the findings. In some cases, teachers were not aware of their scores until they asked the principal for a copy to include in their portfolios or to use during job interviews (Anonymous District Teacher, personal communication, January 20, 2016).

The lack of differentiation in ratings for teachers resulted in districts failing to reward effective teachers or dismiss ineffective teachers. As pointed out in the 2009 Widget Effect report, principals admitted that they had teachers in their schools who they knew were not meeting expectations although they received satisfactory ratings. In fact,

81% of school principals and 57% of teachers stated that there were tenured teachers in their schools with poor performance, yet half of the districts studied in the Widget Effect failed to dismiss a single teacher over a period of two to five years for poor performance. Honestly, failing to document the struggles of teachers and the support provided to them to facilitate improvement made it virtually impossible to exit teachers who were doing students a disservice.

To address these issues, the Obama Administration initiated Race to the Top, as part of the American Recovery and Reinvestment Act of 2009 (U.S. Department of Education, 2009). To earn funding from the national government, states had to demonstrate that they were making changes in key areas that would impact student outcomes. Modifying teacher evaluation systems to incorporate multiple measures and align support to teacher development was a key area of focus for not only the national government, but also for many districts. Since 2009, nineteen states have received Race to the Top funding for creating plans to address the key education reform areas. Tennessee was one of the first states to receive a Race to the Top award. With the award of $500 million, the state was required to implement its education reform plan over a four-year period. The state's plan involved adopting key education reform components, such as Common Core, and revamping the state's educator evaluation model. Moving away from the traditional teacher evaluation system where the only component was classroom observations, the state of Tennessee revised its teacher rubric and included measures of student achievement and growth in teacher evaluation ratings. When NAEP

results were released in 2014, it revealed some of the fruit of Tennessee's labor. Tennessee students were identified as the fastest improving in the nation (Camera, 2015).

Instead of adopting the Tennessee Educator Acceleration Model, which is used in most districts across the state, the district developed an alternate model of teacher effectiveness. The Teacher Effectiveness Measure (TEM) used in a large, urban district in the southeastern United States, is not only comprised of classroom observation ratings and student achievement outcomes, but also incorporates additional variables associated with student learning outcomes. The state of Tennessee made a huge step in requiring multiple observations for all teachers each year and included an achievement component to hold teachers accountable for student performance. The Measures of Effective Teaching project, a national research projected lead by Tom Kane and funded by the Gates Foundation, has shown, however, that combining observations of practice, student achievement, and student feedback into teacher evaluations increases the reliability of results even more. In fact, student feedback measures, such as Tripod surveys, have proven to be stronger predictors of teacher performance than traditional measures like degree attainment (Partee, 2012).

Although value-added scores face criticism across the country, by design, they are intended to level the playing field for teachers and students. Instead of holding teachers accountable for absolute student achievement, they are held accountable for the student growth that they facilitate each year. Even if a student is performing below grade-level, the classroom teacher is not penalized. Teachers are expected to help students meet

minimum growth expectations set by the state. Essentially, a student should grow an academic year for every year of instruction instead of losing ground.

While value-added scores serve as a semi-control for the background characteristics of students that may impact achievement, observation ratings do not take into account extraneous variables that may impact classroom interactions. In fact, research has shown that teachers in high-poverty schools often receive observation ratings that are substantially different from their peers in more affluent schools (Jiang & Sporte, 2016). Since the majority of educators in high-poverty schools are minorities, a concern is raised as to whether evaluation ratings are influenced by school characteristics or if they reflect the actual performance of these teachers when compared to their peers.

Purpose of the Study

The purpose of this study is to determine whether components of the TEM evaluation system provide an accurate assessment of teacher effectiveness, and if so does that hold for all teachers. The TEM is a multiple measure model comprised of five components that should paint a picture of a teacher's effectiveness. Although a teacher's scores will not be the same across all components, there should be alignment between what evaluators see in classrooms, how students perceive the instruction that they receive, and the actual gains that students make in core content areas.

The TEM model should serve as a tool for district and school leaders to assess teacher performance and align development opportunities to evaluative results. It is essential that the system serve as a valid measure of effectiveness for all teachers. Results of various studies have revealed that teacher evaluation scores may be influenced by

teacher and school-level demographics. To better understand the relationship between teacher evaluation ratings and school-level characteristics, this study also examined evaluation scores for teachers in schools with varying levels of Non-White, ELL, SPED, and Economically Disadvantaged students. At the teacher-level, the study sought to determine whether there are differences in the ratings of teachers with varying levels of experience and from different ethnic backgrounds.

Definition of Terms

Throughout this study, terms are used that are linked to student achievement nationally or are specific to the teacher evaluation efforts in Tennessee and the district. While general definitions of the terms are below, further knowledge of many of these terms is built throughout this study:

- *NAEP* is the National Assessment of Educational Progress. Often described as our nation's "yardstick", it has been used since 1969 to assess what students across the U.S. can do. National comparisons are based on the math and reading results of 4^{th} and 8^{th} grade students (NCES, 2015).
- *Tennessee Educator Acceleration Model (TEAM)* - Evaluation model for Tennessee teachers that was first implemented in the 2011-12 school year (Tennessee Department of Education, 2012).
- *Teacher Effectiveness Measure (TEM)* is the evaluation model currently used in Shelby County Schools. This evaluation model was developed by district staff, as part of the Teacher Effectiveness Initiative. The model is comprised of several components with varied weightings: value-added, 35%; student achievement,

15%; classroom observations, 35%; stakeholder surveys, 5%; and content knowledge, 10%.

- *TVAAS* - The Tennessee Value-Added Assessment System estimates student growth in relation to prior achievement on standardized assessments. Teachers' scores are not based on absolute achievement of students. Instead, scores are derived based on how much students are predicted to grow based on prior performance and the actual growth that a teacher facilitates (Magouirk, 2014).
- *TEM Rubric* - The TEM rubric is used to assess the instructional practices of teachers. Teachers are rated based on their performance in the following domains: Plan, Teach, Reflect and Adjust, and Cultivate Learning Environment.
- *Classroom Observations* - Classroom observations are based on the TEM rubric. These observations are typically conducted by principals, assistant principals, district staff, and instructional coaches with administrative licensure.
- *TEM General Education Rubric* - The TEM General Education Rubric is used to assess the instructional practices of teachers in general education classrooms. It is not used to assess librarians, SPED teachers, coaches, or guidance counselors.
- *Tripod Surveys* - TRIPOD surveys are a component of the TEM model, currently in use in Shelby County Schools. These surveys are used to assess students' perceptions of their teachers' classroom practices. These surveys do not measure how well students like their teachers. Survey results provide teachers with invaluable feedback that can be used to strengthen their instructional practice (Ferguson, 2015).

- *Construct* -These are the seven components of effective teaching included in Tripod surveys. Research has shown links between these constructs and student engagement and achievement: Care, Captivate, Consolidate, Classroom Management (formerly Control), Confer, Clarify, and Challenge (Ferguson, 2015).
- *ELL* - This represents English Language Learners, or students who are not native English speakers (National Council of Teachers of English, 2008).
- *SPED* - This represents Special Education students, who have special academic needs because of physical disability, learning disability, or behavior problems (Special Education Guide, 2016).
- *Economically Disadvantaged* - This represents the percentage of students at each school who qualify for free and reduced price lunch.
- *American Recovery and Reinvestment Act (2009)* - A plan signed by President Obama to improve the quality of American life through various measures, including expanding educational opportunities (Department of the Treasury, 2015).
- *Race to the Top* - A $4 billion grant designed by the federal government to incentivize states to design and implement education reform policies to improve student outcomes (U. S. Department of Education, 2010).
- *Novice teacher* - A teacher with less than five years of teaching experience (Kim, 2011).

- *Veteran teacher* - A teacher with five or more years of teaching experience (Kim, 2011).
- *Instructional Culture Insight Index Score* - Is a percentile score that compares instructional culture across schools. The Insight Score is based on teacher responses to key survey items that measure school instructional culture (TNTP, 2012).

Conceptual Framework

The study examines the issue of teacher effectiveness through the lens of social systems theory. Social systems theory considers the various parts, or systems, that are a part of the school, as well as the interaction between and among the parts (Hanson, 1973). An educational system is both the process and outcome of the relationships among its components (teachers, leaders, curriculum and content, students, and climate and culture and the relationship this system has with its environment (King & Frick, 1999). One problem that has historically been true within the cyclical reform efforts, is that change has been made in factions rather than comprehensively. For example, systems theory tells us that when a therapist treats the additive parent and helps them reach sobriety but fails to also work with the broader family systems, the desired impact is not realized. Likewise, tinkering with only one component of the educational system will minimize the impact of that reform. Therefore, as investments have been made in the area of teacher effectiveness, it is important to understand the efficacy of the various models in order to better understand the role of teacher effectiveness within the broader social system of education.

Additionally, the lack of student achievement has pushed the educational field, particularly urban education, into continual reform. And there have been several iterations of reform movement. Under the current reforms, principals are expected to be strong instructional leaders who are charged with enhancing the academic achievement and outcomes for all students. There has been a realization that in part, because of the sheer volume of the problem, principals are not able to drive the scale and pace of change needed alone. Principals are therefore having to pull teachers and their effectiveness into the accountability puzzle. If teachers are going to be involved in being accountable for enhancing student achievement, there must be reliable ways to measure their effectiveness.

Martin Haberman (1995) coined the phrase start teachers and talks about the role that these teachers play in enhancing the academic achievement for students, particularly those in poverty. According to Haberman, teachers who are or are likely to be successful with students exhibit certain traits. These traits tell a story about how these star teachers, as a part of their role within the broader social system of education are able to interact with students, with one another, and within the broader school context. Many of these same skills identified by Haberman (1995) are those that are examined in numerous teacher evaluation models, including the TEM. It is imperative to understand these skills and how to measure them, particularly within the context of the teachers' role as an aspect of the broader accountability puzzle.

Research Questions and Hypotheses

The following research questions were examined in this study:

Primary Question 1. Are the components of the TEM an accurate assessment of teacher performance? And if so, does that hold consistently across teacher demographic characteristics of race and years of experience?

Secondary Question 1. What is the relationship between multiple measures of teacher effectiveness: teacher observation scores, teacher-level student growth scores, and student perceptions of teacher performance?

Primary Hypothesis 1

$H1_0$: There is no relationship between teacher observation scores, TVAAS scores, and Tripod scores.

H1a: There is a relationship between teacher observation scores, TVAAS scores, and Tripod scores.

Secondary Question 2. How are teacher effectiveness ratings distributed for teachers of different races and teachers with varying levels of experience?

Subquestion 1a. What is the distribution of teacher observation scores for White teachers compared to their Non-White peers and Novice teachers compared to Veteran teachers?

Subquestion 1b. What is the distribution of Tripod scores for White teachers compared to their Non-White peers and Novice teachers compared to Veteran teachers?

***Subquestion 1c*.** What is the distribution of TVAAS scores for White teachers compared to their Non-White peers and Novice teachers compared to Veteran teachers?

Secondary Hypothesis 1

$H1a_0$: There is no difference between the distribution of observation, Tripod, and TVAAS scores for teachers for teachers of different races and with varying years of experience.

H1aa: There is a difference between the distribution of observation, Tripod, and TVAAS scores for teachers of different races and with varying years of experience.

Primary Question 2. Do characteristics of teachers and schools predict teacher evaluation ratings?

Secondary Question 3. Do characteristics of teachers and schools predict the observation scores assigned to teachers?

***Subquestion 3a*.** How are observation scores influenced by school-level characteristics (% Economically Disadvantaged, ELL, SPED, Non-White students, School Culture)?

***Subquestion 3b*.** How are observation scores influenced by teacher characteristics (Race, Years of Experience)?

$H3_0$: There is not a predictive relationship between characteristics of teachers and schools and the observation scores assigned to teachers.

H3a: There is a predictive relationship between characteristics of teachers and schools and the observation scores assigned to teachers.

Sub-hypothesis $H3a_0$: There is not a predictive relationship between school characteristics and the observation scores assigned to teachers.

Sub-hypothesis $H3a_a$: There is a predictive relationship between school characteristics and the observation scores assigned to teachers.

Sub-hypothesis $H3b_0$: There is not a predictive relationship between teacher characteristics and the observation scores assigned to teachers.

Sub-hypothesis $H3b_a$: There is a predictive relationship between teacher characteristics and the observation scores assigned to teachers.

Secondary Question 4. Do characteristics of teachers and schools predict the Tripod scores assigned to teachers?

Subquestion 4a. How are Tripod scores influenced by school-level characteristics (% Economically Disadvantaged, ELL, SPED, Non-White students, School Culture)?

Subquestion 4b. How are Tripod scores influenced by teacher characteristics (Race, Years of Experience)?

$H4_0$: There is not a predictive relationship between characteristics of teachers and schools and the Tripod scores assigned to teachers.

H4a: There is a predictive relationship between characteristics of teachers and schools and the Tripod scores assigned to teachers.

Sub-hypothesis $H4a_0$: There is not a predictive relationship between school characteristics and the Tripod scores assigned to teachers.

Sub-hypothesis $H4a_a$: There is a predictive relationship between school characteristics and the Tripod scores assigned to teachers.

Sub-hypothesis $H4b_0$: There is not a predictive relationship between teacher characteristics and the Tripod scores assigned to teachers.

Sub-hypothesis $H4b_a$: There is a predictive relationship between teacher characteristics and the Tripod scores assigned to teachers.

Secondary Question 5. Do characteristics of teachers and schools predict the Individual TVAAS scores assigned to teachers?

Subquestion 5a. How are Individual TVAAS scores influenced by school-level characteristics (% Economically Disadvantaged, ELL, SPED, Non-White students, School Culture)?

Subquestion 5b. How are Individual TVAAS scores influenced by teacher characteristics (Race, Years of Experience)?

$H5_0$: There is not a predictive relationship between characteristics of teachers and schools and the Individual TVAAS scores assigned to teachers.

H5a: There is a predictive relationship between characteristics of teachers and schools and the Individual TVAAS scores assigned to teachers.

Sub-hypothesis $H5a_0$: There is not a predictive relationship between school characteristics and the Individual TVAAS scores assigned to teachers.

Sub-hypothesis $H5a_a$: There is a predictive relationship between school characteristics and the Individual TVAAS scores assigned to teachers.

Sub-hypothesis $H5b_0$: There is not a predictive relationship between teacher characteristics and the Individual TVAAS scores assigned to teachers.

Sub-hypothesis $H5b_a$: There is a predictive relationship between teacher characteristics and the Individual TVAAS scores assigned to teachers.

Significance of Study

For years, teachers and school leaders were not held accountable for student outcomes. The lack of accountability resulted in students leaving high school ill-prepared for college level coursework. Thus, too many students were required to enroll in remedial coursework. Enrolling in remedial coursework not only impacted students during the first

year of college but had more far-reaching impact. Enrollment in remedial coursework is correlated with graduation rates. The majority of students who enroll in remedial courses never receive a college degree (SREB, 2010). It is imperative that American students leave high school better prepared for the rigor of college-level courses.

Since research has shown that a teacher has a greater impact on student outcomes than any other school-related factor, teacher evaluations must provide an accurate assessment of teacher performance. Accurate assessments of performance will allow school districts to align development and support efforts to identified areas of need for individual teachers. Over time, districts will be able to track whether development efforts are effective in improving teacher practice. More importantly, accurate evaluations will allow school leaders to determine which teachers at their schools are improving. They will also be able to determine whether teachers are improving at the desired rate. When teachers are not improving over time, however, school leaders will be able to use the results of evaluations and evidence of the support provided to exit teachers from their schools. While this may sound harsh, every year with an ineffective teacher reduces the likelihood that students will experience success in college and career.

This study examined the extent to which the TEM is providing an accurate assessment of teacher performance. It also revealed whether the evaluation model accurately assesses performance for teachers despite the type of school that they work in. If evaluation ratings for teachers are dependent on teacher and school characteristics, this research study will lay the foundation for district and state officials to reexamine the way teachers are being evaluated.

Limitations

This study had some limitations that may impact the generalizability of its results. Teacher performance data were collected for the 12-13 school year. This study was a snapshot of one year versus a year-over-year analysis. Teachers may and many do, grow over time. Therefore, it is important to contextualize these findings within the limitation that it represents one moment in a more longitudinal picture of a teacher's overarching professional career. Additionally, there have been numerous changes in the structure of the district, professional development platforms, coaching models, and observation rubrics since that time. Therefore, it is possible that some of the findings may be different if data from a more recent school year were utilized.

It is also assumed that the Tripod constructs represent effective teaching practices for all teachers and are predictive of student achievement for all students. It may be the case that Tripod survey results are not predictive of achievement for students in Shelby County Schools or that the predictive power may differ for students from different socioeconomic backgrounds.

Another limitation of this study was the fact that only results of one evaluation system were examined. Since evaluation systems across the nation are comprised of different measures, this study's results may not be generalizable to all school districts.

Study Overview

This research study is comprised of five chapters. Chapter 1 serves as the blueprint for the entire study. This chapter provided background into our issue in a national and local context. Additionally, this chapter revealed why it is important to

examine our research questions in light of current issues. Chapter 2 provides a review of literature as it relates to: (a) teacher effectiveness, (b) teacher evaluation, (c) characteristics of teachers in low and high poverty schools, and (d) the impact of effective teachers on student outcomes. The conceptual framework used within this research study is also further outlined in Chapter 2. Methodology and statistical procedures are outlined in Chapter 3. This includes a breakdown of how the data were collected and analyzed. While findings will be outlined at a high-level in Chapter 4, the culmination of the entire research study is Chapter 5, where the findings will be synthesized. This synthesis will include recommendations for the evaluative process for teachers as well as additional topics for exploration.

Chapter 2: Literature Review

Accountability

The war against educational disparity had its start during the presidency of Lyndon Johnson. The signing of the original Elementary and Secondary Education Act of 1965 was linked to Johnson's "War on Poverty" and envisioned as the tool to transform education for many American students (The Social Welfare History Project, 2014). This act ensured that the federal government allocated more than $1 billion in additional funding to schools serving large proportions of economically disadvantaged students. It was believed that the passage of ESEA would eliminate the effects of poverty on economically disadvantaged students and dramatically increase the rate at which these students earned high school diplomas (Klein, 2015).

Over the last 50 years, the original ESEA has been revisited and amended many times to ensure states and local school districts are adequately addressing the needs of all students. Within a decade, Title I evolved from Johnson's initial plan to spend funds directly on low-income students to the development of schoolwide programs to address the needs of all students in high poverty schools (Klein, 2015).

During the Clinton presidency, ESEA was reauthorized by the Improving America's Schools Act, IASA. Passing this law was a victory not only for the Clinton administration but also for economically disadvantaged students. It required teachers and school leaders to hold economically disadvantaged students to the same standards as their peers (Mead, 2007). This law also required states to identify schools that needed additional intervention to facilitate improvement.

IASA failed to have the intended impact on academic outcomes for students. As a result, the federal government passed a major reauthorization in 2001, No Child Left Behind (Skinner, 2009). Based on the No Child Left Behind Act (NCLB), all students were expected to achieve proficiency in mathematics and reading by the 2013-2014 school year. To that aim, schools were required to make Adequate Yearly Progress, AYP, each school year to receive a passing grade (Mathis, 2006). AYP was based on whether students met specific targets in mathematics and reading each school year. In addition to focusing on the performance of students overall, AYP was also dependent on how various subgroups, such as low-income, racial minorities, and SPED students performed (Johnson, 2007).

In addition to new accountability standards for schools, NCLB placed stricter standards on teachers. Teachers were held accountable for whether their students attained minimum scores for proficiency on state assessments. NCLB did not consider how students were performing when they entered a teacher's classroom (Turner, 2015). In fact, educators in urban schools were expected to facilitate students in reaching the same proficiency levels as students in suburban areas. NCLB disadvantaged teachers in high poverty schools where students often entered their classrooms one or more years below grade level. Even if a teacher facilitated a student in growing several grade levels during an academic year, they still received a failing score, by NCLB standards, if their students did not meet grade level expectations.

The Nation's Report Card, which is based on the NAEP assessment, shows how fourth and eighth grade students are performing nationally. Based on a representative

sample of students, our nation's students were performing very poorly in 2003, the year after the No Child Left Behind Act was instituted. Less than 1/3 of the nation's fourth and eighth grade students achieved proficiency in Reading. While close to 1/3 of fourth grade students were proficient in math, only 29% of eighth grade students were proficient as measured by NAEP (Wirt et al., 2003). NCLB was established to improve statistics such as these.

Over time, the demands of NCLB became increasingly unpopular with educators and teachers' unions. In many states, students were over-tested. This resulted in more time being spent "teaching to the test" which resulted in less time to teach the skills that students needed to be successful in post-secondary ventures (Walker, 2015). NCLB also became unpopular because of the transitions and restructuring that occurred in schools when they failed to meet AYP. Schools that failed to meet AYP for six years were often taken over by their state department of education and required to replace most of their teaching staff. Even after such drastic changes, many of these schools faced challenges that prevented all students from reaching proficiency in math and reading.

Although NCLB required vast changes for teachers and schools, it did not deliver on its promise of proficiency for all students. In fact, over the course of 11 years and several reauthorizations of ESEA, there was only a 7-point increase in the percent of fourth and eighth grade students identified as proficient and advanced by the NAEP reading assessment (National Center for Education Statistics, 2013). Although NAEP results from 1990 revealed that students struggled more in Mathematics than in Reading,

mathematics improvement by 2013 eclipsed that for Reading. Approximately 42% of fourth graders and 36% of eighth graders were proficient in Mathematics.

The failure of NCLB to truly transform academic outcomes and the realization that a one size fits all model would not facilitate improvement for all American students led to the 2015 reauthorization of ESEA. The Every Student Succeeds Act, ESSA, replaced the NCLB Act of 2001 (Korte, 2015). In August, ESEA waivers will expire and ESSA will go into effect. One of the major changes that will occur because of ESSA is a shift in accountability from the federal government to states and districts. Districts and states will have the autonomy to create their own accountability systems, as well as identify strategies to improve student performance (Camera, 2015).

There are key areas that ESSA requires districts and states to focus on. States must establish long-term goals for improving outcomes for all students and students in specific subgroups (The Education Trust, 2016). Although the federal government has provided states and districts with some parameters for what these goals should include, they won't influence the goals that are set (Korte, 2015). Goals should reflect a focus on state assessments and graduation rates, while taking into account the improvement specific subgroups must make to narrow the achievement gap (The Education Trust, 2016).

ESSA also requires states to develop a process for assigning performance ratings to schools. These ratings will be comprised of various components including proficiency on state assessments, proficiency for English Learners, and other measures of academic performance. States must also identify Title I schools that are underperforming. The

federal government has defined several measures that should be used to determine which schools are underperforming. Schools performing in the bottom 5% on state tests, graduating less than 67% of students, and failing to meet the needs of student subgroups should receive additional intervention from their district and state leaders (National Conference of State Legislatures, 2015). Unlike NCLB, ESSA does not outline specific interventions that districts and states must implement to improve chronically underperforming schools.

The emergence of ESSA is also changing many of the educator requirements set by NCLB. The NCLB act required teachers to be highly-qualified in the subject areas that they taught. The highly qualified requirement essentially meant that teachers had to hold a degree and certification in the subject area that they taught (Sawchuk, 2016). While this requirement did prevent schools from staffing classrooms with teachers who failed to meet certain standards, it didn't guarantee that these teachers would facilitate better student outcomes than individuals without the highly qualified status. Under ESSA, the federal government will afford states the opportunity to define a highly qualified or effective teacher within their state context. States will be required to share data on teacher qualifications, experience, and fields of study, as well as the distribution of these teachers in low and high poverty schools (Connally & Tooley, 2016). States must ensure that the distribution of effective teachers is equitable.

In addition to removing regulations for how effective or highly qualified teachers are defined, the national government will allow states to determine how teachers will be evaluated. The flexibility given to districts under ESSA is far removed from the

requirements of NCLB waivers. States receiving NCLB waivers had to ensure that a substantial percentage of teacher evaluation scores were comprised of a student growth measure. States will not be required to link teacher performance evaluations to standardized test scores. Federal funding can be used to sustain these practices, however. Districts will also have the flexibility to totally revamp their evaluation systems when ESEA waivers expire (Ravitch, 2016).

Achievement Gap

The first Elementary and Secondary Education Act was implemented 50 years ago. Since then, ESEA has been reauthorized seven times. More federal dollars have been invested in ESEA than for any other education initiative. Although there have been significant investments by the federal government into ESEA and many of its programs, such as Title I, student performance is still subpar. Overall, students in the United States are lagging behind many industrialized nations academically. Additionally, the achievement gaps between subgroups of students have failed to close over five decades.

To assess the achievement of our nation's students and their preparedness for success in college and careers, states administer standardized tests to assess student performance. Ideally, assessment data should provide parents with accurate information on their child's performance so they can provide them with the appropriate supports at home. A report released by Achieve in 2015 revealed that this is not always the case. States have administered assessments to students across many grade levels each year, however, these assessments were not painting the most accurate pictures of student performance (Achieve, 2015). For more than half of states across the nation, there was a

30-point discrepancy between the 2013 NAEP and their state assessments. Essentially, state assessment results communicated the idea that students were meeting expectations, while NAEP results were showing that students are not prepared for success in more rigorous coursework, college, and careers. This "honesty gap", the difference between the results of state assessments and NAEP, prevented parents from being able to work with educators to improve the performance of their children.

Over the last five years, most states have either adopted the Common Core State Standards or refined their existing standards to reflect the demands that students will face in college. In addition to adopting rigorous standards, states have also refined the assessments used to measure student learning. The new report released by Achieve revealed that the changes enacted by many states have resulted in a narrowing of the "honesty gap" (2016). Based on the current report, 16 schools have eliminated their honesty gap or narrowed the gap to 5 percentage points. There are still substantial gaps between NAEP results and state assessment scores in Tennessee, however. The 8th grade mathematics gap widened resulting in a 25-point discrepancy between NAEP results and state assessment scores (Balakit, 2016).

According to NAEP results released this year, average scores earned by high school students changed slightly from 2013-2015. Reading and math scores dropped by one point (Hinckley, 2016). Results also revealed a drop in college preparedness rates, as measured by NAEP. Only 37% of high school students are prepared for college level reading and math courses (Zernike, 2016). In addition to subpar college preparedness rates for all students, NAEP also revealed that the gap between the achievement of

struggling students and their high-performing peers widened. On the reading assessment, performance for students in the Bottom 10% and the bottom 25% have declined while scores for high performing students have improved. NAEP results for fourth and eighth grade students were not promising either. Math scores dropped for students in both grade levels. Reading scores remained stagnant with slightly more than 1/3 of fourth grade and eighth grade students achieving proficiency (The Nation's Report Card, 2015).

The lack of improvement in reading and math proficiency over time is not the most troubling statistic highlighted by student assessment data. The persistent achievement gap is even more troubling. Thinking back to the beginning of ESEA in 1965, one of the aims was improving performance of students from high poverty backgrounds. Even after many changes to the education landscape and several reauthorizations of ESEA, achievement gaps persist between racial and socioeconomic subgroups.

As a mandate of the Civil Rights Act of 1964, a study to access the quality of schools attended by Black and White students was conducted (Hanushek, 2016). Data from more than 300 schools and 600,000 students were analyzed and synthesized to produce the report, "Equality of Education Opportunity", in 1966. This report, which is often termed the Coleman Report, highlighted the differences in performance of Black and White students. The report revealed a stark contrast between the achievement of Black and White students in both math and reading. At the time of the report, the average Black 12th grader performed at the 13th percentile of the scoring distribution for White

students (Hanushek, 2016). In other words, 87% of White students outperformed the average Black 12th grader.

A recent study examined five years of student assessment data in districts that have completed a decade of federal and state accountability initiatives focused on closing the achievement gap between Black students and their peers (Sparks, 2016). Racial achievement gaps existed in all of the 2,500 school districts studied except Detroit. Researchers contend that this is the case since both White and Black students in Detroit are performing poorly.

An analysis of 2013 NAEP results revealed that racial achievement gaps persist 50 years after the release of the Coleman Report. White students are still significantly outperforming their Black peers. Achievement of the average Black 12th grader lags behind 81% of their peers in math and 78% of their peers in reading (Hanushek, 2016). To put the lack of improvement in perspective, Hanushek asserted that it will take "two and a half centuries before the black-white math gap closes and over one and a half centuries until the reading gap closes" if the current pace of improvement continues.

For decades, a gap has existed between the performance of students from poor families and their affluent peers. The Elementary and Secondary Education Act of 1965 was partially motivated by the differences in performance between poor students and their peers from less disadvantaged backgrounds. Some researchers even assert that the effect of poverty on achievement is so substantial that it can stand alone in impacting academic success. (Biddle & Payne, 2012). While the impact of poverty cannot be denied, factors associated with poverty may contribute just as much to the achievement

gap. The achievement gap between low and high SES students may be exacerbated by limited resources and exposure to less effective teachers (Burris & Heubert, 2006). Students who reside in low income homes often lack access to the same materials and resources as students from affluent backgrounds (Lubienski & Lubienski, 2005). Additionally, students from low socioeconomic backgrounds often attend schools with a high concentration of students from poverty-stricken backgrounds. Most districts face challenges when trying to attract the most highly qualified teachers to these schools.

A recent study examined the academic achievement of students in the nation's 100 largest school districts (Rich, 2016). Data collected from tests administered to third through eighth graders over a four-year period showed that the performance of students in school districts with the highest concentrations of poverty trailed the performance of their peers in the richest districts by four grade levels.

A 2013 ACT report highlighted differences between low SES students and their peers. When students from low socioeconomic backgrounds were asked about their educational aspirations, their responses were pretty similar to students from more affluent backgrounds. More than 80% of low income students stated that they planned to attend college after graduating from high school. Their performance on the college readiness assessment, however, indicated that they are ill-prepared for the demands of college coursework. Only 27% of students in households with incomes less than $36,000 met the college readiness benchmarks in Reading while 24% of these students met the Math benchmark. Students from affluent backgrounds fared much better. When students'

household incomes exceeded $100,000, 64% met the college readiness benchmark in Reading and 66% met the benchmark in Math (Klein, 2015).

Results from the Educational Longitudinal Study of 2002 highlighted differences in educational attainment for students from different backgrounds (Musu-Gillette, 2015). A group of 10th grade students were surveyed in 2002 and throughout their postsecondary ventures. When the educational attainment levels of these students were assessed ten years later, several interesting findings were revealed. Only 14% of students from low socioeconomic backgrounds earned Bachelor's degrees compared to 60% of students from high socioeconomic backgrounds. Even high performing students from low socioeconomic backgrounds failed to attain the same levels of performance as their peers. Only 40% of low-income students performing in the top quartile for math during their sophomore year earned a Bachelor's degree. Contrastingly, 74% of high performing students from affluent backgrounds earned a college degree (NCES, 2014).

Teacher Effectiveness

The academic performance of students across the nation is impacted by an array of factors. Family characteristics, such as socioeconomic status, parental education level, and the number of parents in a household all impact academic achievement. Student outcomes are also influenced considerably by the extent to which parents are involved in their child's education. Aware of the impact of parental engagement on student performance, school districts have encouraged parents for years to become actively engaged in schools. A research study conducted using data from the National Educational Longitudinal Study (NELS) revealed that schools would have to spend $1,000 more per

student to impact student performance to the same degree as parental involvement (Houtenville, 2008).

While student background, family characteristics, and parental engagement impact student achievement, school leaders and districts cannot control these factors. Educational institutions can control the quality of teachers that students are exposed to, however. Since a teacher is the single most important school-related factor in regard to student outcomes, it is imperative that all students have access to effective teachers.

Although a teacher is the school-related factor that has the greatest impact on student outcomes, a consistent definition of teacher effectiveness does not exist. Over the last decade, many states and school districts have shifted their focus to the impact of teacher quality on student performance, however, there isn't real clarity around what defines an effective teacher or how effective teaching should be measured in most states. For this reason, state departments of educations and school districts across the country have incorporated various models into their evaluation systems to fully understand the measures that define effective teaching.

It is important to note the distinction between teacher effectiveness and teacher quality. Effectiveness, a term under the umbrella of accountability, is a construct that the educational field has in recent years, adopted from the field of business. Effectiveness, as borrowed as a business term connotes efficiency, or return on investment. Anderson (1991) defined teacher effectiveness as the extent to which a teacher consistently achieves goals that focus on student learning. By its origins, this definition squarely puts the emphasis on a numerical or metric-based approached to measuring and articulating

teacher effectiveness. This framework is different from the more traditional education approaches to considering a teacher's performance, which is grounded in language of teacher quality and competence, which is the extent to which a teacher has the requisite knowledge and skills that enable him or her to behave a certain way during the actual process of teaching. The notion of examining inputs as a measure of teacher quality differs from a more outcomes-based approach that is used to measure teacher effectiveness (Dunkin, 1997). For the purpose of this study, the latter definition of teacher performance, grounded in the ability to deliver results for students is used to consider the research questions surrounding teacher effectiveness.

As one of the trailblazers of the changes to the teacher evaluation landscape, Shelby County Schools in 2013 outlined teacher effectiveness as:

> The individual impact that the teacher has on student learning and achievement documented using both quantitative data (e.g., value added data, student test scores, individual and group performance assessment of students, student projects, and student classroom performance) and qualitative data (e.g., peer observations and surveys, observational information, or teacher professionalism, when applicable, from teachers, district curriculum specialists, principals, parents, and students).

In addition to outlining how teacher effectiveness would be assessed, the district highlighted that effective teachers would be those who facilitated a year of student growth each academic school year and fostered improved academic outcomes for students from diverse backgrounds.

It is important to note at this point, that these measures do not come without surrounding controversy. Although value-added scores face criticism across the country, by design, they are intended to level the playing field for teachers and students. Instead of holding teachers accountable for absolute student achievement, they are held accountable for the student growth that they facilitate each year. Even if a student is performing below grade-level, the classroom teacher is not penalized. Teachers are expected to help students meet minimum growth expectations set by the state. Essentially, a student should grow an academic year for every year of instruction instead of losing ground.

Previous studies have also examined the influence of teacher sorting and its unintended effects on the ability to accurately measure teacher effectiveness (Kalogrides, Loeb, & Beteille, 2012). The researchers from Stanford found that a process of sorting and assignment of teachers within schools results in the most effective teachers not being matched with the students who may in fact need them the most. Additionally, findings from the research by Kalogrides et al. (2012) has implications that influence how the estimation of teacher value-added scores are calculated. Most value-added models assume that the process of assigning students with teachers is not critical to the calculation of estimates. While it is not random, the calculation of value added scores is treated as if the assignments should not be controlled for or factored in the calculation. However, some research suggests that teacher assignment is dependent upon a host of factors and therefore would in fact matter when calculating value added scores (Kalogrides et al, 2012).

While value-added scores serve as a semi-control for the background characteristics of students that may impact achievement, observation ratings do not take into account extraneous variables that may impact classroom interactions. In fact, research has shown that teachers in high-poverty schools often receive observation ratings that are substantially different from their peers in more affluent schools (Jiang & Sporte, 2016). Since the majority of educators in high-poverty schools are minorities, a concern is raised as to whether evaluation ratings are influenced by school characteristics or if they reflect the actual performance of these teachers when compared to their peers. The analysis by the Brookings Institution (2016) revealed that nearly all the opportunities for improvement to teacher evaluation systems are in the area of classroom observations rather than in test score gains.

And although district definitions and measurement of teacher effectiveness may differ, there is one constant. Observations of instructional practice remain the foundation for assessing teacher effectiveness across the nation. An analysis of evaluation data from four districts highlighted the reliance of school districts on observation data. The analysis by the Brookings Institution (2014) revealed that nearly all the opportunities for improvement to teacher evaluation systems are in the area of classroom observations rather than in test score gains. This does not mean that value-added scores are not accurate predictors of teacher effectiveness. Instead, it highlights the fact that the foundation for defining and measuring teacher effectiveness hinges largely on observations of instructional practice because these scores are available for most teachers.

On the other hand, only teachers of specific courses and grade levels receive growth scores.

Since observations of practice are the metric that most districts rely on to measure teacher quality, it is imperative that the tools and processes utilized are valid and reliable. *Fixing Classroom Observations*, a report by TNTP, a national non-profit organization, revealed several issues with classroom observations that need to be addressed to facilitate their use as improvement tools for teachers (TNTP, 2013). The report highlighted the need to streamline observation rubrics. Evaluators are not able to provide teachers with useful feedback if they are required to assess performance on too many metrics. This includes having evaluators to rate teachers on every indicator on a rubric and on characteristics that cannot be observed in the classroom like professionalism. Instead, districts need to identify indicators that are most aligned to student outcomes and rate on these, as well as simplifying the rubrics that evaluators are required to use.

There are additional issues that should be addressed to ensure classroom observations facilitate improvement in teacher practice. Districts must shift away from a compliance mindset where most of the focus is on whether observations were completed on time and the score that teachers received (Dewitt, 2013). Instead, districts need systems that ensure the feedback that evaluators provide to teachers is specific and actionable. Districts must work to ensure that the observation process focuses more on providing teachers with effective feedback instead of compliance with district policies (TNTP, 2013).

For decades, classroom observations served as the sole or primary measure of teacher practice. Research has shown, however, that classroom observations, when used alone, are not an accurate predictor of student outcomes. In fact, research conducted by the MET Project revealed that even when teachers are observed four times during a school year by multiple observers, the accuracy of the ratings are lower than for any of the multiple measure models that they tested (MET Project, 2013). Since a single measure of performance cannot define a teacher's effectiveness, many districts and states have reformed their teacher evaluation systems to improve their ability to identify low performing teachers, provide actionable feedback, individualize support, and retain high performers (The New Teacher Project, 2010).

As research by various organizations started to reveal the importance of utilizing multiple measures to accurately assess teacher performance, some districts began to explore the use of measures of students' perceptions of their learning environments and teacher practices. Although student surveys have been used as a measure of performance in postsecondary educational institutions for years, student perceptions were not included as a measure of performance for K-12 teachers until recently (Hanover Research, 2013). In many cases, it was believed that students were not equipped to provide reliable responses about teacher performance.

There is a huge disparity between classroom observations, that have always been used as a measure of teacher effectiveness, and student perceptions, however. Classroom observations are based on an administrator visiting a classroom a few times a school year, whereas students are in classrooms for approximately 180 days. The time that students

are exposed to a teacher's instruction gives them the best lens into the quality of the instruction that is being delivered (MET Project, 2012). As a result, student surveys are being used in some districts to assess teacher actions and the learning environments that they foster.

In the 2012 report, "Asking Students about Teaching", the predictive validity of the Tripod survey, a measure of student perceptions, was highlighted. Authors of the report revealed that a teachers' survey results were predictive of their student achievement scores. In fact, teachers who earned top quartile Tripod scores facilitated an additional 4.6 months of learning gains in mathematics (2012). Although the relationship between Tripod scores and ELA results was not as significant, the correlation was still positive.

In addition to qualitative measures, quantitative measures are also a component in many multiple measure evaluation systems. Multiple measure models often include measures of student learning as well as student growth over time as measured by state tests. Combining achievement and growth scores with teacher observations and student surveys further increases the reliability and the predictive power of evaluation ratings (Partee, 2012).

Multiple measure models serve another purpose in addition to increasing the accuracy of teacher ratings. They afford school leaders the opportunity to provide teachers with timely, actionable feedback. Student achievement tests are typically administered during the spring of each school year. By this time, students have been exposed to a teacher's instructional practice, whether effective or ineffective, for seven

months. Testing timelines and the receipt of results prevent teachers from being able to adjust their instruction based on areas of deficiency. Even when schools receive these results, they fail to provide detailed information that can be used to improve teacher performance the next school year. Contrastingly, classroom observations and student surveys have the potential to provide results that can be used to drive feedback conversations and teacher development activities that lead to improvement in teacher quality. And although these measures are not imperfect, when combined they are seemingly able to provide a more accurate picture of a teacher's overall effectiveness (MET Project, 2010).

Teacher Demographics and Teacher Effectiveness

It is important to call out the role and importance of teacher demographic characteristics, particularly years of experience and race, within the teacher effectiveness discussion. States and districts have an obligation to ensure that observation tools are not biased toward particular groups of teachers. During their study, the Brookings Institution (2014) identified bias in the way scores were assigned to teachers. Teachers who taught higher performing students received higher observation ratings, while teachers with lower performing students received lower observation ratings. To further exacerbate this issue, a study by Chaplin, Gill, Thompkins, and Miller (2014) revealed that teachers with a higher proportion of low income students in their classes also received lower observation ratings than their peers. Consequently, these issues could disadvantage two particular groups of teachers. Teachers who serve in the most high-need schools may receive lower observation ratings than their peers who serve students in more affluent areas. Less

experienced teachers could also be impacted by bias in observation instruments. Studies have shown that the lowest performing students are assigned to novice teachers more often than effective teachers with more experience (Mead, 2012). The impact of disproportionately assigning the lowest performing students to novice teachers is two-fold. These teachers receive lower observation ratings than their more experienced peers, but more importantly, students with subpar performance in the past are not exposed to the high-quality instructional practices that will lead to improvements in their educational performance.

In other research, this pattern has held and remains troubling when looking at teacher assignments and distributions. Researchers at Stanford found that certain teachers, often those with less experience, those who had attended less-competitive colleges, female teachers, as well as teachers of color, were more likely to be found working in lower performing schools. Additionally, these teachers were found to be most often working with lower-performing students, when compared to other teachers within the same school (Kalogrides et. al., 2012). Teachers have preferences when it comes to the types of students that they teach. Many studies have found that when given a choice, teachers prefer to teach in schools with easier to serve, higher-performing student populations (Kalogrides et al., 2012). However, the researchers found that teachers who were minority were more often found to be assigned to students of color and students who were lower performing (Kalogrides et al., 2012). As this research has suggested, if the sorting of teachers across and within schools is both influenced by and influences student outcomes; then it is imperative that teacher evaluation models take into account

this previous research. The research examining the efficacy of the models should seek to ask and answer questions about the influence of these teacher demographics in accurately measuring teacher effectiveness to ensure fairness, validity and reliability of teacher evaluation models.

School Culture

As described earlier, social systems theory considers the various parts, or systems, that are a part of the school, as well as the interaction between and among the parts (Hanson, 1973). An educational system is both the process and outcome of the relationships among its components (teachers, leaders, curriculum and content, students, and climate and culture and the relationship this system has with its environment (King & Frick, 1999). Therefore, there is a logical inference that the interplay of these components creates a particular culture and instructional culture specific to a school. The question then becomes whether or not school culture plays an important role in influencing the success of a school. Studies have found that school culture does in fact play an important role in student achievement (Hanushek, 1997). School culture is defined as the way teachers and other adults in the school work together, as well as the set of values and beliefs that they have in common. A positive school climate and school culture promote students' ability to learn (ASCD, 2017). Therefore, it is important to understand the role that the school's culture does or does not play in influencing the ability of a teacher to deliver at least a year of academic growth per year of instruction, put another way, to be effective. If in fact school culture plays a significant role in enabling a teacher's

effectiveness, it is important to understand what the implication is for measuring individual teacher effectiveness across and within schools.

In a 2012 study by TNTP, researchers found that affirmative responses from three particular survey questions had the strongest correlation to retention of effective teachers and higher student achievement in reading and math. These three questions were utilized to create an instructional culture score, or the Insight index for schools as a measure for school culture (TNTP, 2012). The three questions were: "Teachers at my school share a common vision of what effective teaching looks like; The expectations for effective teaching are clearly defined at my school; and My school is committed to improving my instructional practice" (TNTP, 2012). The instructional culture score, or Insight index, will serve as a proxy measure of school culture in the study.

Chapter 3: Methodology

Introduction

While it is important to acknowledge that there are other home-related variables that influence student achievement, this study was focused on understanding those variables that the school system has agency over such as teacher effectiveness and school culture. The purpose of this study was two-fold and answers two major research questions. First, it determined whether the teacher evaluation model utilized in the district provides an accurate assessment of teacher quality. This study examined various factors including relationships between the different components of the teacher evaluation model. This analysis revealed whether different components of the evaluation model paint similar pictures of educator effectiveness. Second, this study also assessed whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study sought to determine whether the teacher evaluation system accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve. To achieve these goals, classroom observation data, teacher-level student growth scores, and student perceptions data were examined. This chapter is comprised of the methodology and design procedures of the study, the population to be examined, instruments used in the study, and data collection and analysis.

Methodology

This study examined the distribution of effective ratings for teachers on the general education track in traditional public schools in the district. The distributions of

observation and overall evaluation ratings were examined. The study also examined the relationships between the components of the teacher evaluation system in the district to determine whether they are consistent in their assessment of teacher performance. The study also explored whether characteristics of teachers and schools are predictive of teachers' observation scores and overall evaluation ratings.

Research Questions

This study focused on five smaller research questions in order to answer the two major research questions. These research questions were used to the relationships between components of teacher evaluations, and the impact of teachers and school-level characteristics on teacher ratings.

In order to answer the first major research question of whether different components of the evaluation model paint similar pictures of educator effectiveness, the study explored the following research questions:

1. What is the relationship between multiple measures of teacher effectiveness: teacher observation scores, teacher-level student growth scores, and student perceptions of teacher performance?
2. Do the distributions of teacher effectiveness ratings differ for teachers of different races and teachers with varying years of experience?

The second major research question examined in this study assessed whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study sought to determine whether the teacher evaluation system accurately assesses the

performance of all teachers despite their race and the unique characteristics of the schools where they serve. The research subquestions that answer this second major question are:

3. Do characteristics of teachers and schools predict the observation scores assigned to teachers?
4. Do characteristics of teachers and schools predict the Tripod scores assigned to teachers?
5. Do characteristics of teachers and schools predict the individual growth scores assigned to teachers?

Research Design

This study employed three quantitative research designs. These designs are descriptive, correlational, and non-experimental causal-comparative. Quantitative research methodology was utilized in this study for multiple reasons. Quantitative research affords researchers the opportunity to test hypotheses. This is especially important in this study since the research questions and hypotheses being examined show whether the evaluation system used for thousands of teachers in a large urban school district disadvantages teachers of certain races and those who serve the most underprivileged students. Quantitative research also allows for the generalizability of findings and trends from a sample population to a larger population of interest. This method provides school districts that are using a model similar to the TEM and working with similar populations of teachers with the opportunity to learn from the results of this study and apply some of the findings to how they evaluate teacher performance. Quantitative research is also more reliable and objective than some more qualitative

methods. Eliminating most of the subjectivity that is inherent in qualitative research will allow districts to make decisions that are based on clear and compelling evidence (Ramona, 2011).

The descriptive design is the basis of the entire analysis. It is a very simple process that shows what the current state of data is. The descriptive design shows the current distribution of teachers' evaluation ratings. This design not only provides summary statistics and frequencies of evaluation ratings but also calculates measures of central tendency (Cohen, Manion, & Morrison, 2007). Mean scores of groups are compared in the descriptive design phase of the study.

The correlational design is used to test the relationships between the components of the teacher evaluation model. Correlational designs show whether changes in one variable result in changes in another variable. For example, if weight increases as height increases, there is a positive correlation between the two variables. The next step would be determining the magnitude or size of this relationship. Correlations vary in magnitude from -1 to +1. Values that are closer to 1 are the strongest. Values that are closer to zero show that relationships are non-existent or minimal (Simon, 2011). While correlational designs show evidence of relationships, causation cannot be inferred from a correlational design.

The non-experimental, causal-comparative design was used to determine whether the evaluation ratings assigned to teachers are influenced by demographic characteristics. Although a causal-comparative study is not as strong as an experimental research design, it does allow for the examination of causal relationships (Lodico, Spaulding, & Voegtle,

2006). It is especially helpful for trying to understand why differences exist between groups. A causal-comparative study is limited in that it is based on data that have already been collected or differences that have already occurred. As a result, variables cannot be manipulated.

Population and Sample

Teachers from all elementary, middle, and high schools in the district who were assigned to the general education track during the educator evaluation process constituted the population of interest in the present study. Teachers in specialized schools and charter schools were excluded from this study. Educators who specialize in Special Education and Counseling were excluded from this study. The schools included in this study vary considerably in size. The school in this study with the smallest teaching staff had eight teachers, while the school with the largest teaching staff had 73 teachers. The overall distribution analysis was comprised of all teachers in the population of interest. Stratified random sampling was utilized to identify individuals from different races to include in the analysis that tests whether evaluation scores differ for teachers from different racial backgrounds.

A teacher is the most important school-level factor that influences student success. To ensure that students are exposed to optimal learning opportunities, teacher effectiveness must be accurately assessed. To accomplish this goal, teachers in traditional schools with observation data and evaluation data constituted the population of interest in this study. To ensure comparability of effectiveness ratings, only teachers on the general education track were included in the study. In other words, teachers on the Special

Education track were not included in the study. The observation rubric used to evaluate these teachers differs from the rubric used to observe instruction of teachers on the general education track. To ensure observation and evaluation ratings of teachers can be compared, these teachers were not included in the analysis. Although guidance counselors are classified as educators, they were not included in the analysis since they are not responsible for classroom instruction.

This study only included teachers in traditional, public schools within the district. Teachers in charter schools, alternative schools, and specialized schools are not included in the population of interest. Although some charter schools within the district utilized the teacher evaluation model employed in the district, teachers in charter schools were not always required to adhere to the same standards as teachers in traditional district schools. Implementation practices also differed in charter schools; principals did not always complete observations within the same timeframe as teachers in traditional schools.

Teachers in alternative schools are faced with behavioral challenges that their peers in traditional district schools are not always exposed to. This could result in teachers in alternative schools receiving student survey results that are substantially different than their peers in traditional schools because of the student populations that they serve. Specialized schools often serve students who suffer from severe emotional, mental, and/or physical handicaps. As a result, teachers in these schools were not included in the study since the evaluation and observation ratings that they receive would not be comparable to the ratings assigned to teachers in traditional district schools.

During the 2012-13 school year, there were 4,200 teachers on the general education track in traditional district schools who received observation and evaluation ratings through the district's teacher evaluation system. This entire population of teachers was used to examine the districtwide distribution of teacher ratings. The next analysis examined the relationships between observation ratings and teacher-level student growth ratings. For this part of the study, the effectiveness ratings of 1,311 teachers were examined. To further explore relationships between evaluation components, the relationship between classroom observation ratings and student perceptions was explored. For this part of the study, data for 1,786 teachers was examined. Relationships between evaluation components and teacher and school-level characteristics was also examined.

To assess whether differences exist between groups, mean performance ratings of teachers from various subgroups was examined. This includes examining the Tripod, Observation, and Individual TVAAS scores of novice vs. veteran teachers and white vs. non-white teachers.

Since this study aimed to examine whether factors, such as race, are predictive of teacher effectiveness ratings, 1,072 teachers without demographic data were then eliminated from the pool of general education teachers in the study. From this reduced population of 3,127 teachers, 1,055 White teachers and 1,055 Non-White teachers were randomly selected for inclusion in the observation and evaluation analysis. Utilizing samples that are similar in size will ensure results are comparable and representative of the population of teachers.

Instruments

The Teacher Effectiveness Measure (TEM) General Education Observation Rubric was utilized to assess the quality of classroom practices. The TEM rubric was initially implemented in classrooms across the districts in the 2011-2012 school year. The version of the TEM Rubric used in this study assesses teacher practice in four key domains: Plan, Teach, Cultivate Learning Environment, and Reflect and Adjust. Observations are rated on a scale of 1-5. A rating of 1 is equal to "Significantly Below Expectations" while a rating of 5 is equal to "Significantly Above Expectations".

TRIPOD surveys were utilized to measure students' perceptions of teacher practice. TRIPOD surveys measure teacher practice in seven key areas. These areas, or constructs, are linked to student engagement and performance on standardized assessments. The seven TRIPOD constructs included in teachers' overall ratings in this study are: Care, Captivate, Clarify, Challenge, Consolidate, Confer, and Control. The Control construct has recently been changed to Classroom Management. Tripod surveys are rated on a scale of 1-5. A rating of 1 is equal to "Significantly Below Expectations" while a rating of 5 is equal to "Significantly Above Expectations".

Teacher-level student growth scores were derived from the Tennessee Value-Added Assessment System (TVAAS). These growth scores, or value-added scores, measure the impact that teachers have on the academic gains of students. Value-added scores measure the amount of progress a teachers' students made from one school year to the next. Value-added scores for individual teachers will be used in this study. Value-added scores are assigned on a scale of 1-5. A rating of 1 is equal to "Least Effective"

while a rating of 5 is equal to “Most Effective”. A rating of 1 is assigned to teachers whose students made significantly less progress than the Growth Standard, while a rating of 5 is assigned to teachers whose students made significantly more progress than the Growth Standard. Variable descriptions are provided in Table 1.

Data Collection

For the purposes of this study, pre-existing data was used. Approval to use historical teacher evaluation data from the district’s TEM system was granted from the district’s research department. Initially, employee background and demographic data were collected from the district. This includes unique identification numbers for teachers, hire data, schools, and race. De-identified, matched educator evaluation data for the 2011-2012 and 2012-2013 school years were also collected from the district. This includes classroom observation data, TRIPOD student survey data, and TVAAS growth data. To assess whether school-level characteristics influence performance ratings, aggregate data on ELL, SPED, and Economically Disadvantaged students was collected. Additionally, data from the 2013-2014 Instructional Culture Insight Survey served as a proxy measure of school culture.

Table 1

Variable Definitions

Variables	Definitions
Race	Race is coded 0 = White; 1 = Non-White.
Years of Experience	The number of years that a teacher has taught in the district.
Novice	A teacher with three or fewer years of classroom experience
Veteran	A teacher with more than three years of classroom experience
Observation Score	Component in the TEM Model comprised of teachers' instructional practice ratings. Observation scores range from 1.00-5.00
Observation Rating	The rating that is input into a teacher's evaluation. This rating is based on the observation score that ranges from 1.00-5.00. A teacher is assigned one of the five observation ratings: • *1- Performing Significantly Below Expectations* • *2- Performing Below Expectations* • *3- Meeting Expectations* • *4- Performing Above Expectations* • *5- Performing Significantly Above Expectations*
Individual TVAAS Index	Score based on the ratio of a teacher's growth effect score to the standard error.

Table 1 (Continued)

Variables	Definitions
Individual TVAAS Level	The TVAAS Level assigned to a teacher based on their index score. TVAAS levels are assigned based on the following ranges: • *1- Index Scores < -2* • *2- Index Scores of -2 to -1* • *3- Index Scores of -1 to 1* • *4- Index Scores of 1 to 2* • *5- Index Scores > 2*
TRIPOD Score	A variable representing the average favorability rating a teacher received across each of the survey constructs. Favorability ratings are based on the percent of students in agreement on survey items.
TRIPOD Level	The rating assigned to a teacher based on their overall TRIPOD score. The level is assigned based on the quintile of a teacher's TRIPOD Score. Levels are based on the following ranges: • *1- 1-20%* • *2- 21-40%* • *3- 41-60%* • *4- 61-80%* • *5- 81-100%*
FRPL Rate	This is a proxy for the socioeconomic level of a school that represents the percentage of students who qualify for free and reduced price lunch.

Table 1 (Continued)

Variables	Definitions
ELL Rate	This rate represents the percentage of students identified as Limited English Proficient.
SPED Rate	This rate represents the percentage of students identified as special education or students with disabilities.
Instructional Culture Insight Score	This is a percentile score that compares instructional culture across schools. The Insight Score is based on specific responses to survey items that measure school instructional culture.

Data Analysis

The initial research questions in this study required an examination of the rating distributions for classroom observations, student surveys, and growth scores. To answer these questions, frequencies of the three variables were calculated.

To ascertain the magnitude and direction of the relationships between each component of the TEM, Pearson's product moment correlation coefficients were calculated. Each variable in the correlation analysis was measured at the interval or ratio level. This means that each variable included in the correlation analysis must be continuous. A scatterplot was examined to ensure that linearity exists between the variables.

To determine whether characteristics of teachers and schools predict the evaluation ratings assigned to teachers, a linear regression analysis was conducted. Linear regression requires that four assumptions be met. The first assumption is that the relationship between variables is linear. The second assumption is that multivariate normality exists. Examination of a histogram was used to determine whether normality exists. The third assumption is that the variables are independent and multicollinearity does not exist. Tolerance and variance inflation factors were examined to ensure this assumption is met. A boxplot was used to determine whether any outliers exist. The fourth assumption is homoscedasticity. Homoscedasticity requires the error terms across all independent variables to be the same. A scatterplot was used to ensure the homoscedasticity assumption is met.

The Statistical Package for the Social Sciences (SPSS) was used to analyze the data. Descriptive statistics were run to get a summary of the frequencies of each evaluation rating for the components in the TEM model. This analysis showed how evaluation scores were distributed across the entire district. Additionally, schools were placed into quartiles based on the percentage of students who qualify for free and reduced price lunch. Descriptive statistics were run on each quartile of schools to see the distribution of teacher ratings. Pearson's R were also calculated to assess the magnitude and direction of the relationships between the components in the TEM model.

Independent samples t-tests was used to determine whether the mean observation, TVAAS, and Tripod scores differ for teachers in schools with varying levels of students receiving free and reduced priced lunch. T-tests were also used to determine whether mean scores differ for White and Non-White educators. Finally, mean scores were examined to determine whether there are significant differences between the scores of novice and veteran teachers.

To determine whether teacher and school level characteristics are predictors of teacher evaluation ratings, three linear regression models were estimated. The first model explored whether teacher race and the percent of students receiving free and reduced priced lunch are predictors of teachers' observation scores. The second model explored whether teacher race and the percent of students receiving free and reduced priced lunch are predictors of teachers' overall Tripod scores. The third model explored whether teacher race and the percent of students receiving free and reduced priced lunch are predictors of teachers' individual TVAAS scores. Additional variables, including years of

experience, the percent of ELL and SPED students, and Insight scores that measure school culture were included in the models.

Summary

There are several purposes for this chapter. This chapter first outlined the research methodology for this study. The chapter then described which research questions this study attempted to address, as well as the research designs that was implemented. The population of interest in the study, the instruments under examination, and the variables to be included in the study's analysis were then described. Finally, the chapter outlines in detail which analyses would be conducted to answer the research questions.

Chapter 4: Results

This study focused on five smaller research questions in order to answer the two major research questions. These research questions were used to the relationships between components of teacher evaluations, and the impact of teachers and school-level characteristics on teacher ratings.

In order to answer the first major research question of whether different components of the evaluation model paint similar pictures of educator effectiveness, the study explored the following research questions:

1. What is the relationship between multiple measures of teacher effectiveness: teacher observation scores, teacher-level student growth scores, and student perceptions of teacher performance?
2. Do the distributions of teacher effectiveness ratings vary for teachers based on race or years of experience, or who are teaching in different school-level demographic contexts?

The second major research question examined in this study assessed whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study sought to determine whether the teacher evaluation system accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve. The research subquestions that answer this second major question are:

3. Do characteristics of teachers and schools predict the observation scores assigned to teachers?

4. Do characteristics of teachers and schools predict the Tripod scores assigned to teachers?
5. Do characteristics of teachers and schools predict the individual growth scores assigned to teachers?

Data on Teacher Effectiveness Ratings

Initially, descriptive statistics of teacher observation ratings were examined The analysis sought to determine how teacher effectiveness ratings were distributed. Table 2 is comprised of the means, standard deviations, and correlations for the entire population of educators used in the first phase of this study. Observation ratings, individual growth scores, and Tripod scores are based on scales of 1-5. An examination of classroom observation scores revealed that the majority of teachers were identified as Performing Above Expectations and Performing Significantly Above Expectations. Of the 4,200 educators with observation scores, 97% (n = 4,087) received ratings of Effective or higher. On a 5-point scale, the average observation score for all educators was 4.06.

Of the 4,200 educators in the entire population of teachers with observation ratings, there were 1,786 with Tripod scores. An examination of the ratings revealed that 82% (n = 1,471) were rated as Effective or higher. On a 5-point scale, the average Tripod score was .67. On average, 67% of teachers' students responded favorably to Tripod constructs around classroom culture and experiences.

Of the 4,200 educators in the entire population of teachers with observation ratings, there were 1,311 who taught core classes and received individual growth scores. An examination of these ratings revealed that 66% (n = 867) of the teachers received

individual growth scores of three or higher. Growth scores of three or higher represent teachers who are growing students at least one year for a year of instruction. Additionally, 41% of teachers in the study are providing students with more than a year of growth for each year of instruction. On a 5-point scale, the average individual growth score was .5937.

Findings Based on the Relationships Among Multiple Measures of Teacher Effectiveness

The first major research question was whether different components of the TEM paint similar pictures of teacher effectiveness, and if so, does the similarity hold consistently across teacher demographic characteristics of race and years of experience.

To answer the first major question, the subsequently described analyses were conducted. Relationships between teacher observation scores, TVAAS scores, and Tripod scores were also examined. Calculation of correlation coefficients indicated that the null hypothesis should be rejected. Significant relationships existed between observation, growth, and Tripod scores. There was a moderate, positive correlation between observation scores and TVAAS scores, which demonstrates that teachers' growth scores improved as their observation scores improved. This is promising since observation scores should paint a similar picture of teacher effectiveness as growth scores. Although the magnitude of the relationship was smaller than the relationship between observation and TVAAS scores, Tripod scores were also positively correlated with observation ratings. As teachers' observation ratings increased, Tripod scores also increased. TVAAS

scores were also significantly correlated with Tripod scores. Teachers with higher TVAAS scores also received higher Tripod scores.

Pearson's r was calculated to answer the first research question in the study. This part of the analysis sought to assess the relationships between multiple measures of teacher effectiveness and demographic variables. Observation scores, individual growth scores, Tripod scores, teacher-level characteristics, and school-level characteristics were included in the correlation analysis. Each of these metrics was measured at the interval level. Average observation scores were significantly correlated with all variables in the model. Observation scores were most highly correlated with individual growth scores. As observation scores increased, individual growth scores also increased (.359, $p < .01$). Teachers with higher observation scores also received higher Tripod scores (.118, $p <$. 01). Additionally, teachers with more years of teaching experience and those who worked in schools with higher Instructional Culture Insight scores (.162, $p < .01$) received higher observation ratings. The relationship between observation scores and school-level FRPL status (-.264, $p < .01$) was negative. As the percent of students in a school receiving free and reduced priced lunch increased, teachers' observation scores decreased.

Tripod scores were also significantly correlated with all variables in the model. As Tripod scores increased, individual growth scores also increased (.129, $p <.01$).

Table 2

Correlations, Means, and Standard Deviations of Variables in Descriptive Analysis

Variables	1	2	3	4	5	6
1. Years of Experience	____					
2. FRPL Rate	-.071***	____				
3. Observation Score	.083***	-.264***	____			
4. Individual TVAAS Index	-.051	-.113***	.359***	____		
5. Tripod Score	.055	.153***	.118***	.129*	____	
6. Instructional Culture Insight Score	.073***	-.071***	.162 ***	.138***	.159***	____
Means	11.5261	85.2808	4.058	.5937	.6687	.4724
Standard Deviations	7.8356	15.7409	.6200	4.4064	.1635	.2665

** $p < .01$. *** $p < .001$

Additionally, teachers with more years of teaching experience (.055, $p < .05$) and those who worked in schools with higher Instructional Culture Insight scores (.159, $p < .01$) received higher Tripod scores. While observation scores were negatively correlated with FRPL status, this was not the case for Tripod scores. The correlation between Tripod scores and FRPL status (.153, $p < .01$) was statistically significant and positive. As the percent of students in a school receiving free and reduced priced lunch increased, teachers' Tripod scores also increased.

Individual growth scores were not significantly correlated with all variables in the model. The relationship between individual growth scores and years of experience was not significant. Individual growth scores were significantly correlated with Instructional Culture Insight scores (.138, $p < .01$) and FRPL status (-.113, $p < .01$). Teachers who worked in schools with higher Instructional Culture Insight scores earned higher individual growth scores, while teachers who worked in schools serving a larger population of economically disadvantaged students earned lower individual growth scores.

To better understand how teachers in high-poverty schools are performing compared to their peers in more affluent schools, schools were placed into quartiles based on the percentage of students receiving free and reduced priced lunch. Table 3 outlines the percentage of teachers who were rated as Effective or higher for each evaluation component. An examination of observation scores revealed that more than 95% of teachers in bottom quartile and top quartile schools were rated as Effective or higher.

Table 3

Percentage of Teachers Rated as Effective or Higher by School Quartile

	Bottom Quartile of FRPL	*Middle Quartiles of FRPL*	*Top Quartile of FRPL*
Evaluation Component			
Observation Score	98.57% (n=1537)	97% (n=1968)	95.4% (n=695)
Individual TVAAS Index	66.6% (n=503)	65.52% (n=609)	66.83% (n=199)
Tripod Score	79.46% (n=560)	81.97% (n=915)	88.75% (n=311)

Although the percentage of teachers receiving Effective or higher individual growth scores did not differ much across FRPL quartiles, fewer teachers were identified as meeting expectations compared to the results from the observation analysis. There was more variance in the percentage of teachers receiving Tripod scores of Effective or higher. Teachers in schools with the highest rates of poverty received the highest ratings more often than teachers in schools serving less disadvantaged students.

Before conducting the next phase of analysis, stratified random sampling was employed. Comparable samples of White (n = 1,055) and Non-White (n = 1,055) teachers were selected. independent samples t-tests were used to determine whether the mean observation, Tripod, and individual growth scores differed for groups of teachers. Testing for the assumption of equality of variances revealed that the data failed to meet this assumption. SPSS automatically corrects for failing to meet this assumption. The reported statistics are based on the data for equal variances not assumed.

The first set of independent samples t-tests was conducted to compare mean observation, Tripod, and individual growth scores for White and Non-White teachers. Table 4 is comprised of the results of the independent samples *t* Tests for White and Non-White teachers. The test for observation scores was found to be statistically significant, $t(2047) = 11.418$, $p < .01$; $d = .497$. The effect size for this analysis approximated Cohen's (1988) convention for a medium effect (d=.50). These results indicate that White teachers ($M = 4.2135$, $SD = .60451$) received observation scores that were significantly higher than the scores received by Non-White teachers ($M = 3.8831$, $SD = .71963$). The

Table 4

Results of Independent Samples t Tests and Descriptive Statistics for Teacher Performance by Race

Outcome	Group						95% CI for Mean Difference	t
	White			Non-White				
	M	SD	n	M	SD	N		
Observation Scores	4.214	.604	1055	3.883	.720	1055	.274, .387	11.418***
Tripod Scores	.663	.168	490	.637	.148	573	.007, .045	2.676**
Individual TVAAS Index	.846	.118	252	.599	4.592	858	-.349, .842	.814

** $p < .01$. *** $p < .001$

test for Tripod scores was also found to be statistically significant, $t(984) = 2.676$, $p < .01$; $d = .165$. These results indicate that White teachers ($M = .663$, $SD = .168$) received Tripod scores that were significantly higher than the scores received by Non-White teachers ($M = .637$, $SD = .148$).

The test for individual growth scores was not found to be statistically significant. Although White teachers earned growth scores that were higher than Non-White teachers, the difference in the average across groups was not significant. The next set of independent samples t-tests explored differences in evaluation scores for novice and veteran teachers. Novice teachers have three or fewer years of experience, while veteran teachers have more than three years of experience. Testing for the assumption of equality of variances revealed that the data failed to meet this assumption for observation and Tripod data. SPSS automatically corrects for failing to meet this assumption. The reported statistics for these dependent variables were based on data for equal variances not assumed.

Table 5 is comprised of the results of the independent samples t-tests for Novice and Veteran teachers. The first independent samples t-test was conducted to compare mean observation scores for Novice and Veteran teachers. The test for observation scores was not found to be statistically significant. Although the mean score for Veteran teachers was higher than the mean score for Novice teachers, the difference in the average across groups was not significant.

Table 5

Results of Independent Samples t Tests and Descriptive Statistics for Teacher Performance by Experience Level

Outcome	Group						95% CI for Mean Difference	t
	Novice			Veteran				
	M	SD	n	M	SD	n		
Observation Scores	4.055	.614	248	4.076	.678	1665	-.104, .063	-.485
Tripod Scores	.604	.131	107	.655	.163	860	-.078, -.026	-3.686***
Individual TVAAS Index	1.677	4.835	124	.386	4.434	826	.381, 2.201	2.802**

** $p < .01$. *** $p < .001$

The test for Tripod scores was found to be statistically significant, $t(150) = -3.686$, $p < .01$; $d = .34$. These results indicate that Veteran teachers ($M = .655$, $SD = .163$) received Tripod scores that were significantly higher than the scores received by Novice teachers ($M = .604$, $SD = .131$).

The test for individual growth scores was found to be statistically significant, $t(155) = 2.802$, $p < .01$; $d = .28$. These results indicate that Veteran teachers ($M = .385$, $SD = 4.84$) received growth scores that were significantly lower than the scores received by Novice teachers ($M = 1.6768$, $SD = 4.43$).

Findings Based on the Influence of Teacher and School-level Characteristics on Teacher Effectiveness Ratings

The second major research question was whether school and teacher characteristics are predictive of education evaluation ratings. The next phase of analysis focused on answering this question. The regression analyses conducted in this study explored whether characteristics of teachers and schools predicted teachers' observation, Tripod, and growth scores. The first regression model was estimated to test the null hypothesis that there was not a predictive relationship between characteristics of teachers and schools and the observation scores assigned to teachers. Before conducting the analysis, tests were run to see if statistical assumptions were met. The histogram of standardized residuals indicated that the data contained approximately normally distributed errors, as did the normal P-P plot of standardized residuals, which showed points that were not completely on the line, but close. The scatterplot of standardized residuals revealed that the data met the assumptions of homogeneity of

variance and linearity. The six independent variables were entered into the regression equation simultaneously. An analysis of standardized residuals was carried out on the data to identify any outliers, which indicated that 14 cases had to be removed. Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Observation Scores: Maximum VIF = 1.384). The data did not meet the assumption of independent errors (Durbin-Watson value < .001).

Table 6 details the results of the full regression analysis. A significant regression equation was found which resulted in rejection of the null hypothesis ($F(6,1279)= 47.594$, $p<.000$) with an R^2 of .183. A significant amount of variance is being explained by the set of independent variables with three of the six variables having significant unique influence on the dependent variable. In order of importance, they were FRPL (β = -.297), Insight score (β = .203), and race (β =-.119). In the presence of the other variables in the model, teachers' observation scores were lower in schools that had higher percentages of students receiving free and reduced priced lunch. Teachers' observation scores were also lower in schools with lower Instructional Culture Insight scores. Additionally, White teachers received observation scores that were significantly higher than the scores received by their Non-White peers. Although the effect of race (β =-.119) succeeded the effects of the aforementioned variables, its effect on observation scores is substantively important. White teachers received higher observation scores than their Non-White peers when other variables in the model are controlled for. In the presence of the other variables in the model, teacher experience, the percentage of SPED students,

Table 6

Results of Regression of Independent Variables on Observation Scores

Independent Variables	*b*	*β*	*t*
Race	-.151	-.119	-4.418***
Years of Experience	.003	.038	1.475
FRPL Rate	-.010	-.297	-9.971***
Instructional Culture Insight Score	.492	.203	7.581***
Percent of ELL Students	.001	.027	.929
Percent of SPED Students	-.002	-.014	-.556

R^2 = .183 (*N*= 1,286, ** $p < .01$; *** $p < .001$)

and the percentage of ELL students in a school did not significantly influence observation scores.

The second regression model was estimated to test the null hypothesis that there is not a predictive relationship between characteristics of teachers and schools and the Tripod scores assigned to teachers. Before conducting the analysis, tests were run to see if statistical assumptions were met. The histogram of standardized residuals indicated that the data contained approximately normally distributed errors, as did the normal P-P plot of standardized residuals, which showed points that were not completely on the line, but close. The scatterplot of standardized residuals revealed that the data met the assumptions of homogeneity of variance and linearity. The six independent variables were entered into the regression equation simultaneously. An analysis of standardized residuals revealed that there were no issues with outliers. Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Tripod Scores: Maximum VIF = 1.421). The data met the assumption of independent errors (Durbin-Watson value = .193).

Table 7 details the results of the full regression analysis. A significant regression equation was found which resulted in rejection of the null hypothesis ($F(6,625) = 15.794$, $p<.000$) with an R^2 of .132. A significant amount of variance is being explained by the set of independent variables with two of the six variables having significant unique influence on the dependent variable. The percentage of ELL students wielded the greatest influence on Tripod scores ($\beta = .238$), while school-level Insight scores wielded the second greatest influence on the dependent variable ($\beta = .174$). In the presence of the other variables in

Table 7

Results of Regression of Independent Variables on Tripod Scores

Independent Variables	*b*	*β*	*t*
Race	-.020	-.066	-1.671
Years of Experience	.001	.066	1.732
FRPL Rate	.001	.094	2.113
Instructional Culture Insight Score	.099	.174	4.359***
Percent of ELL Students	.003	.238	5.498***
Percent of SPED Students	.001	.036	.950

R^2 = .183 (*N*=632, ** $p < .01$; *** $p < .001$)

the model, Tripod scores were higher in schools that had higher percentages of ELL students. Teachers in schools with higher Insight scores also received higher Tripod scores.

The third regression model was estimated to test the null hypothesis that there is not a predictive relationship between characteristics of teachers and schools and the Individual TVAAS scores assigned to teachers. Before conducting the analysis, tests were run to see if statistical assumptions were met. The histogram of standardized residuals indicated that the data contained approximately normally distributed errors, as did the normal P-P plot of standardized residuals, which showed points that were not completely on the line, but close. The scatterplot of standardized residuals revealed that the data met the assumptions of homogeneity of variance and linearity. The six independent variables were entered into the regression equation simultaneously. An analysis of standardized residuals was carried out on the data to identify any outliers, which indicated that four cases had to be removed. Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Individual TVAAS Scores: Maximum VIF = 1.474). The data did not meet the assumption of independent errors (Durbin-Watson value = .002). Table 8 details the results of the full regression analysis. A significant regression equation was found which resulted in rejection of the null hypothesis ($F(6,592) = 7.049$, $p<.000$) with an R^2 of .067. A significant amount of variance is being explained by the set of independent variables with two of the six variables having significant unique influence on the dependent variable. FRPL rate wielded the greatest influence on growth scores ($\beta = -.223$), while school-level

Insight scores wielded the second greatest influence on the dependent variable (β = .128). In the presence of the other variables in the model, growth scores were higher in schools serving students from higher socioeconomic backgrounds. Teachers in schools with higher Insight scores also received higher growth scores.

Table 8

Results of Regression of Independent Variables on Individual TVAAS Scores

Independent Variables	*b*	*β*	*t*
Race	.081	.010	.223
Years of Experience	-.048	-.080	-1.989
FRPL Rate	-.049	-.223	-4.633***
Instructional Culture Insight Score	2.021	.128	3.110**
Percent of ELL Students	.019	.055	1.217
Percent of SPED Students	-.024	-.022	-.524

R^2 = .067 (*N*=603, ** $p < .01$; *** $p < .001$)

Chapter 5: Discussion, Recommendations and Conclusion

Summary

Although the classroom teacher is the most important school-related factor influencing student achievement, districts and schools across the nation have struggled to ensure every child has access to effective teachers. Historically, the teacher evaluation process has been implemented inconsistently in many districts. In fact, there are districts where teachers have historically only been evaluated at five year intervals. In districts where teacher evaluations have been conducted on a more regular basis, the process often consisted of administrators visiting teachers' classrooms 1- 2 times during an academic year. These classroom visits were rarely coupled with actionable feedback which further limited teacher performance and opportunities for improvement.

These systems for evaluating teachers were ineffective at best. Teachers continued to receive evaluation ratings that indicated they were meeting students' needs. Student performance data told a different story, however. Even when teachers were receiving the highest observation ratings, their students were often failing to meet minimum expectations on state assessments. Although school administrators admitted that there were struggling teachers in their schools, for years, many districts failed to dismiss any teachers because of performance. Since evaluation ratings indicated that these teachers were meeting or exceeding expectations, school leaders were essentially devoid of any power to remove ineffective teachers from classrooms. Consequently, students were consistently exposed to ineffective teaching practices.

While these issues affected students at schools across districts, they were even more prevalent in schools that served students from underprivileged backgrounds who were already at risk. High attrition rates often resulted in constant influxes of the least experienced and least effective teachers in schools that served large populations of students from economically disadvantaged backgrounds. Rather than narrowing the achievement gap between economically disadvantaged students and their peers from more affluent backgrounds, issues such as teacher performance further perpetuated the decades long issue.

Over the last 10 years, the federal government began implementing measures to address the challenges with teacher performance that were exacerbating student academic issues and preventing closure of the achievement gap. Initial measures included offering funding to states and districts to motivate them to reform their teacher evaluation systems and provide teachers with the necessary supports to deliver effective instruction to students. Measures of student performance and perceptions were added into teacher evaluation models. These changes represented a shift from traditional, observation-based evaluation systems to multiple measure models intended to hold educators more accountable for student performance.

While many states and districts have implemented multiple measure teacher evaluation models, some of the original challenges have still surfaced. A disproportionate number of teachers are still receiving the highest evaluation ratings. Teacher evaluation ratings and measures of student performance are often painting disparate pictures of teacher effectiveness. Differences between the evaluation ratings of teachers of

economically disadvantaged student populations and teachers of more affluent students often surface also (TNTP, 2013).

To ensure that evaluation metrics can be used to accurately assess teacher performance and drive improvement conversations, schools and districts must ensure that they are unbiased reflections of teacher performance. Teacher evaluation ratings should not be influenced by extraneous variables but should level the playing field for all teachers. Essentially, the ratings obtained through the teacher evaluation process should not be influenced by school or teacher-level characteristics, such as teacher race or the proportion of economically disadvantaged students in a school or classroom.

Ultimately, this research study sought to examine two overarching themes: first, whether the teacher evaluation model utilized in a large, urban district in the southeastern United States provides an accurate assessment of teacher quality and second, whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study was established to determine whether the teacher evaluation system accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve.

Teachers in the district were not observed on a regular basis prior to the reformation of the district's teacher evaluation system. The reformed teacher evaluation model, while not perfect, represented the district's drastic shift away from only observing teachers once every five years to measuring teacher performance every year. This measurement of teacher performance was no longer based solely on observations of classroom instruction but also on metrics such as student performance and perceptions.

Since this shift from past practice represents a clear effort to more accurately assess teacher practice, it is important to examine the extent to which these changes are influencing the way teacher performance is quantified across the district. While classroom observations have been criticized in the past for inaccuracy, measures such as Tripod surveys, which measure students' perceptions of teacher practice, have been upheld as more accurate reflections of teacher performance and better predictors of student performance.

In conducting this research, the researcher had two purposes. The first purpose for conducting this research study was the need to determine whether or not the components of the Teacher Effectiveness Measure (TEM), accurately measure teacher performance. The secondary purpose of this study was examining whether teachers' performance ratings are influenced by teacher and school-level demographics. This study examined whether or not school-related factors such as school culture and the proportion of Non-White, ELL, SPED, and Economically Disadvantaged students influenced teacher ratings. Differences in the ratings of teachers with varying levels of experience and from different ethnic backgrounds were also examined.

Since evaluation components such as student perceptions have been identified as more accurate measures of teacher performance, there should be alignment between teachers' scores on Tripod surveys and the observation ratings that they receive from school and district evaluators.

The teacher evaluation system has been marketed as a tool for district and school leaders to utilize to assess performance, develop and support teachers, and make

employment decisions; therefore, it is essential that it serve as a valid measure of effectiveness for all teachers. Prior to development of the new teacher evaluation system, decisions to retain and dismiss teachers were not strategic. When teacher dismissals did occur, they were rarely based on classroom performance. Similarly, teacher retention decisions were not tied to classroom performance. Since most teachers received the highest performance ratings and were retained each year, teachers who were truly high-performing and pushing the needle on student achievement were not treated differently than their peers.

Shifting to the use of the teacher evaluation system as a tool for decision making necessitates an exploration of its efficacy for evaluating teachers from all racial backgrounds and schools. Since results of previous studies revealed that ratings acquired from some teacher evaluation systems have been influenced by factors other than the level of instruction that teachers deliver, a secondary purpose of this study was examining whether teachers' performance ratings are influenced by teacher and school-level demographics. This study examined whether or not school-related factors such as school culture and the proportion of Non-White, ELL, SPED, and Economically Disadvantaged students influenced teacher ratings. Differences in the ratings of teachers with varying levels of experience and from different ethnic backgrounds were also examined.

Discussion of Findings

This study focused on five smaller research questions in order to answer the two major research questions. These research questions were used to examine the

relationships between components of teacher evaluations, and to examine the impact of teacher and school-level characteristics on teacher ratings.

In order to answer the first major research question of whether different components of the evaluation model paint similar pictures of educator effectiveness, the study explored the following research questions:

1. What is the relationship between multiple measures of teacher effectiveness: teacher observation scores, teacher-level student growth scores, and student perceptions of teacher performance?
2. Do the distributions of teacher effectiveness ratings vary for teachers based on race or years of experience, or who are teaching in different school-level demographic contexts?

The second major research question examined in this study assessed whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study sought to determine whether the teacher evaluation system accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve. The research subquestions that answer this second major question are:

3. Do characteristics of teachers and schools predict the observation scores assigned to teachers?
4. Do characteristics of teachers and schools predict the Tripod scores assigned to teachers?
5. Do characteristics of teachers and schools predict the individual growth scores assigned to teachers?

An analysis of the data revealed the following key findings, discussed in more detail later in the chapter. Related to the first major research question of whether different components of the evaluation model paint similar pictures of educator effectiveness, the key findings were:

- the components of the Teacher Effectiveness Measure (TEM) were correlated, suggesting it is a valid multiple-measures evaluation model;
- despite this, minority teachers were more likely to get lower observation scores than white teachers;
- veteran teachers were more likely to receive higher TRIPOD scores than novice teachers;
- novice teachers were more likely to receive higher TVAAS scores; and
- both observation ratings and TVAAS scores decreased as the proportion of economically disadvantaged students in a school increased.

Related to the second major research question of whether school and teacher characteristics are predictive of educator evaluation ratings, the key findings were:

- the proportion of economically disadvantaged students in schools was the greatest predictor of teacher observation scores. In the presence of the other variables in the model, teachers' observation scores were lower in schools that had higher percentages of students receiving free and reduced priced lunch;
- the second highest predictor of observation scores was the school-level Instructional Culture Insight score. In the presence of other variables in the

regression model, teachers' observation scores were also lower in schools with lower Instructional Culture Insight scores;

- in the presence of other variables in the regression model, White teachers received observation scores that were significantly higher than the scores received by their Non-White peers;
- in the presence of other variables in the second regression model, Tripod scores were higher in schools that had higher percentages of ELL students;
- in the presence of other variables in the second regression model, Teachers in schools with higher Insight scores also received higher Tripod scores;
- in the presence of the other variables in the model, growth scores were higher in schools serving students from higher socioeconomic backgrounds; and
- also, teachers in schools with higher Insight scores also received higher growth scores.

In order to answer the first research question, the researcher examined the correlations between and among the components of the model. The data revealed that the components of the model were correlated. The subquestions of the first research question speak to whether the correlations hold for specific teacher and school demographics. It was interesting that race and years of experience surfaced as significant contextual variables. Relationships between teacher observation scores, TVAAS scores, and Tripod scores were examined. Calculation of correlation coefficients indicated that the null hypothesis should be rejected. Significant relationships existed between observation, growth, and Tripod scores. There was a moderate, positive correlation between

observation scores and TVAAS scores, which demonstrates that teachers' growth scores improved as their observation scores improved. This is promising since observation scores should paint a similar picture of teacher effectiveness as growth scores. Although the magnitude of the relationship was smaller than the relationship between observation and TVAAS scores, Tripod scores were also positively correlated with observation ratings. As teachers' observation ratings increased, Tripod scores also increased. TVAAS scores were also significantly correlated with Tripod scores. Teachers with higher TVAAS scores also received higher Tripod scores.

Also, the distributions for Tripod scores and value-added ratings were more balanced than that for the observation distribution. While 82% of teachers received Effective or Highly Effective Tripod scores, 66% of teachers received TVAAS scores that were Effective or Highly Effective. It is important to note that while 82% of teachers received Effective or Highly Effective Tripod scores when these ratings were transformed to a 5-point scale, the average Tripod score was .67. This score means that 67% of teachers received favorable ratings from their students on Tripod constructs. This demonstrates that the distribution of Tripod scores (67% rated favorably) and TVAAS ratings (66% meeting minimum expectations set by the state) are more similar to each other than to the distribution of observation ratings. The similarity in these distributions is not surprising since research has revealed that Tripod scores are more accurate predictors of performance than other teacher evaluation measures.

In order to answer the secondary question of whether or not ratings differ for teachers of different races and varying years of experience, a subset of ratings for

teachers were examined to determine whether there were differences between these distributions for teachers with varying demographic characteristics. The first variable of interest was teacher race. As a result, independent samples t-tests were conducted to see if statistically significant differences existed between each evaluation component and teacher race. The analysis of race resulted in a partial rejection of the null hypothesis. There were significant differences between the observation and Tripod scores of White and Non-White teachers. For both observation and Tripod scores, White teachers received scores that were significantly higher than their Non-White peers. Compared to the results for the observation analysis, the mean difference between Tripod scores and teacher race (.026) was minimal. There was a difference of .331 in the average observation mean for White and Non-White teachers. Evidence exists to show that minority teachers are more likely to work in higher poverty schools (Belsha, 2016). This statistic held true for this study. Of the sample of White teachers, 44% worked in schools where 90% or more of the students were identified as economically disadvantaged by their school lunch status. Of the sample of Non-White teachers, 70% worked in schools where 90% or more of their students were identified as economically disadvantaged. The greater likelihood of Non-White teachers working in schools that serve more disadvantaged populations may explain some of the differences in the ratings that they received compared to their White peers. This is an important finding to further investigate since retention and dismissal decisions in the district are now predicated on teacher performance ratings and teachers should not be penalized for serving in more challenging schools.

Independent samples t-tests were also employed to examine whether there were differences in evaluation scores for novice and veteran teachers. The results of the analysis result in a partial rejection of the null hypothesis. While observation scores did not differ for novice and veteran teachers, there were significant differences between the Tripod and growth scores of novice and veteran teachers. The finding for Tripod scores was not surprising since additional years of experience are expected to translate into higher levels of performance for teachers. Contrastingly, novice teachers, or teachers with three or fewer years of experience earned growth scores that were significantly higher than the scores that veteran teachers earned. Before discussing this finding, it is important to note the differences in the samples under examination. While there were 124 novice teachers included in the analysis, there were 826 veteran teachers included in the analysis. Barring the difference in sample sizes, this finding raises questions about the growth of veteran teachers. As highlighted in *The Mirage*, researchers found that teachers on average demonstrated growth in the early years, until around year five, and too many teachers plateau before mastering some very critical skills (TNTP, 2015). The significantly higher scores earned by novice teachers could be influenced by a failure of veteran teachers to continuously innovate and learn new strategies to address the needs of an ever-changing student population. Additionally, the difference in ratings could indicate that students who are struggling the most are assigned to the least experienced teachers. If this is the case, these students, who have a greater distance to grow, may make more traction during the school year than their peers who are not multiple grade levels behind.

Additionally, in answering the first research question, further examination of the correlation coefficients revealed interesting relationships between school-level characteristics and observation ratings. The significant, positive relationship between instructional culture and observation ratings raises questions as to whether educators who teach in schools with disproportionate numbers of at-risk students, who often present more behavioral challenges than their peers, are disadvantaged by the observation system. There was a statistically, significant relationship between observation scores and school-level FRPL status. This negative relationship revealed that teachers' observation ratings decreased as the proportion of economically disadvantaged students in schools increased. Individual growth scores, which were significantly correlated with most variables in the model, were also negatively correlated with FRPL rates. Teachers who worked in schools serving a larger population of economically disadvantaged students earned lower individual growth scores. These results raise several concerns. The differences in evaluation ratings may indicate that teachers in high-poverty schools are not delivering the same quality of instruction as their peers in schools with less disadvantaged students. On the other hand, these differences may indicate that it's more difficult for teachers to earn the highest average scores in high-poverty schools due to the unique challenges that these schools face. If this is the case, it raises concerns that teachers who challenge themselves to serve in schools with students who need them the most are being adversely affected by their decision (University of Chicago Consortium of School Research, 2016). The differences in evaluation ratings may result in these teachers

transferring to schools with lower levels of poverty where it appears to be easier to earn higher evaluation ratings.

While Tripod scores were significantly correlated with all variables in the model, the most interesting relationship surfaced between Tripod scores and FRPL status. Unlike the relationship between observation scores and FRPL, the correlation between Tripod scores and FRPL status was positive. Teachers who served in the most disadvantaged schools received the highest Tripod scores. Unlike observation ratings assigned by school leaders, students' perceptions of teacher practice do not seem to be influenced by higher school poverty concentrations.

The next phase of analysis sought to answer the second major research question of whether school and teacher characteristics are predictive of educator evaluation ratings The regression analyses conducted in this study explored whether characteristics of teachers and schools are predictors of observation, Tripod, and growth scores for teachers in the district. To answer subquestion 3, the first regression model was estimated to test the null hypothesis that there was not a predictive relationship between teacher and school-level characteristics and the observation scores assigned to teachers. The results of this analysis resulted in a rejection of the null hypothesis. The proportion of economically disadvantaged students in schools was the greatest predictor of teacher observation scores. A one unit decrease in the FRPL rate yielded a .01 increase in teacher observation ratings. These results could be indicative of less effective teachers in schools that serve the most disadvantaged students. If this is the case, consideration should be given to attracting more effective teachers to schools where students need them most. This could

be accomplished in different ways including offering bonuses to high-quality teachers who are willing to work in these schools, allowing groups of teachers to transfer to these schools together, and implementing strategies to retain high performing teachers. To retain high performing teachers, it is imperative that the district ensures the observation ratings assigned to teachers are accurate and can be used to differentiate teachers who are performing at different levels. On the other hand, the predictive power of FRPL rates could be indicative of issues with the observation system. If teachers' observation ratings are more reflective of the challenges faced in high-poverty schools instead of the level of instruction that teachers are delivering to students, the district must identify ways to level the playing field for teachers in all schools.

The second highest predictor of observation scores was the school-level Instructional Culture Insight score. As can be expected, teachers in schools with higher ratings for school culture received higher observation scores. This finding raises an important task for school leaders. To ensure teachers are best positioned to provide high-quality instruction that meets students' needs, leaders must ensure that their school cultures are conducive to learning and teaching. Since earlier analysis revealed that observation scores are lower in schools that serve larger populations of economically disadvantaged students, it may serve leaders well to intentionally focus on the improvement of school culture. This may help to resolve some of the challenges that teachers experience when working with as-risk groups of students.

This analysis revealed another finding that may warrant additional investigation. Although the effect of race on observation scores was smaller than that of FRPL rates and

Insight scores, it was still a predictor of the observation scores that teachers receive. White teachers received observation scores that were .151 points higher than the scores received by Non-White teachers. As was highlighted earlier in this study, the majority of teachers in high-poverty schools are Non-White. Since teachers in high-poverty schools are receiving lower observation ratings than their peers in less disadvantaged schools, it is expected that the Non-White teachers who serve in these schools are receiving lower ratings than their White peers in schools with lower levels of poverty.

To answer subquestion 4, the second regression model was estimated to test the null hypothesis that there is not a predictive relationship between characteristics of teachers and schools and the Tripod scores assigned to teachers. A significant regression equation was found which resulted in rejection of the null hypothesis. The greatest predictor of Tripod scores was the percentage of ELL students in a school. For every one-unit increase in the percentage of ELL students, there was a .003 increase in teachers' Tripod scores. As part of the Tripod administration process, districts must adhere to certain requirements to address the unique needs of ELL students and the language barriers that they may face. In addition to translating the surveys into students' primary home languages, surveys are also read to students who may have deficiencies that would prevent them from understanding the survey items. School-level Insight scores were also significant predictors of Tripod scores. While Insight scores measure school culture at the school-level, Tripod scores measure various elements of instructional practice, including the extent to which teachers demonstrate that they care for the students in their

classrooms. Since these variables measure similar constructs, it is not surprising that a predictive relationship exists.

To answer subquestion 5, the third regression model was estimated to test the null hypothesis that there is not a predictive relationship between characteristics of teachers and schools and the Individual TVAAS scores assigned to teachers. A significant regression equation was found which resulted in rejection of the null hypothesis. FRPL rate was the greatest predictor of teacher-level student growth scores. The analysis revealed that for every one-unit increase in the percentage of economically disadvantaged students, there was a .048 point decrease in the value-added index score. Unlike measures of proficiency that require all students to meet the same bar to be classified as proficient, value-added scores should partially level the playing field for teachers and students (McCaffrey, 2012). Since the performance of students in high-poverty schools is often lower than the performance of students in schools with lower levels of poverty, failing to meet the state's minimum expectations for growth exacerbates existing issues. If students are not growing, the achievement gap between students in high-poverty schools will never reach the performance of teachers in less disadvantaged schools. Results of this analysis should be further examined to assess factors that are impeding the growth of students from economically disadvantaged backgrounds.

School culture was also a significant predictor of growth scores. For each one-unit increase in the school culture index, there was a 1.931 point increase in growth scores. This finding further emphasizes the importance of focusing on the improvement of school instructional culture. Improving school culture is important for multiple reasons. A

positive school culture affords teachers and school leaders with the opportunity to focus their time on instruction instead of losing invaluable instructional time as a result of an unbalanced focus on behavior issues. Research has shown that school culture also impacts the rate at which effective teachers leave schools which further impacts student performance. In the report, Greenhouse Schools: How Schools Can Build Cultures Where Students and Teachers Thrive, it was revealed that 10% of effective teachers in schools with weak instructional cultures identified school environment and learning culture as the most important reasons for them considering leaving their schools (TNTP, 2012). To improve student performance in low-performing schools, principals must implement strategies to improve school culture so teachers who have the potential to deliver high-quality instruction to students can be retained.

Limitations of Study

While this research study revealed findings that facilitate a better understanding of the multiple-measure teacher evaluation model that was implemented in the district under study during the 2012-2013 school year, there are limitations that could potentially impact the applicability and generalizability of findings. One limitation of this study is the use of historical evaluation data that may not reflect the performance of current teachers in the district or changes that have occurred in the district. Since these data were collected, the district has experienced a merger with the suburban school district and several changes in leadership. Additionally, recent changes in the observation rubric and weightings of components in the evaluation model could potentially influence analysis results and findings. Future research should be conducted to focus on teacher evaluation

data based on the current observation framework and scoring protocols. This research would also be more representative of all teachers in the district.

Additionally, this study was a snapshot of one year versus a year-over-year analysis. Teachers may and many do, grow over time. Therefore, it is important to contextualize these findings within the limitation that it represents one moment in a more longitudinal picture of a teacher's overarching professional career.

While teacher performance levels were available for each of the evaluation components, the analysis in this study required the use of raw data and average scores. The district did not provide raw data for each evaluation component across the requested years. To combat this issue, the data in this study reflect teacher performance across multiple years. While observation and Tripod data from the 2012-2013 school year were utilized, the TVAAS scores used were from the 2011-2012 school year and Insight index scores were from the 2013-2014 school year. Demographic data were also used from the 2013-2014 school year. Future research should incorporate data from the same school year for each evaluation component. This will allow researchers to speak more to the alignment between evaluation components since they will represent all of a teacher's performance during a single school year.

When conducting regression analysis, there are six statistical assumptions that researchers should test for. When the assumptions were tested for this study, data did not meet the assumption of independent errors. Essentially, there was some auto-correlation among the error terms. The Durbin-Watson statistic is used to determine whether this assumption is met. Values of 2 indicate that there is no auto-correlation, whereas values

approaching 0 indicate that there is auto-correlation. For the regression models in this study, Durbin-Watson values ranged from 1.698- 1.913. Future research should examine additional independent variables that may be missing from the model. Since the exclusion of additional evaluation components is limited to how the district's model is defined, there may be additional demographic variables that can be tested to see if they eliminate the auto correlation issues. Since there was some auto-correlation present, an alpha level of .01 was used for all statistical tests instead of the less rigorous significance level of .05.

Recommendations for Future Studies

The data from this study yielded some interesting findings that may warrant additional analyses and be instructive for future research. This study did not use a cohort analysis to examine year over year teacher growth. The teacher evaluation model being used in the district has now been implemented for several years. It would be important to use data for all years of implementation to see if the findings from this study hold and check for internal consistency. By studying the longitudinal results, correlations over time could be analyzed to examine internal consistency. Also, by looking at multiple years of data, which the district does have, it would be possible to explore and gain a deeper understanding of the context and any related implications. Future studies might utilize hierarchical linear modeling, or cluster analysis as a method to analyze multiple years of data.

Relatedly, it would be important to study again in several years to see if teachers who had improved observation scores also had improved TVAAS scores, given the components were correlated. Doing so would indicate that the feedback and professional development that teachers were receiving was meaningful and actually contributing to teacher growth and translating to improved student outcomes.

Additionally, as educators across the country are looking at ways to more meaningfully engage students who have historically struggled, it might be interesting to dig deeper understanding why TRIPOD scores were higher for ELL students in the district. Focus groups with both ELL students and their teachers might yield more qualitative data that would help better understand this finding.

A third possible area for additional research would be to better understand the role of teacher demographics in observation ratings. Future research should probe more into the area to better understand the differences highlighted in this study, and to further explore areas that might arise related to potential bias.

Implications for Practice

Ultimately, this research study sought to examine two overarching themes: whether the teacher evaluation model utilized in the district under study provides an accurate assessment of teacher quality and whether school and teacher characteristics are predictive of educator evaluation ratings. Essentially, the study was established to determine whether the teacher evaluation system accurately assesses the performance of all teachers despite their race and the unique characteristics of the schools where they serve.

While the TEM, the multiple-measure evaluation model utilized in the district under study, is not a perfect model, it does align to recommendations in the MET project. These recommendations were based on research that demonstrated the superiority of the multiple-measure approach above traditional observation-only approaches to evaluating teacher performance. Perfect alignment between components does not exist and is not expected. As revealed by Kane, McCaffrey, Miller, and Staiger (2013) in the final MET report, evaluation model components, when combined, should paint a comprehensive picture of teacher effectiveness. This comprehensive picture is a more accurate reflection of teacher performance than any single evaluation measure provides. Value-added scores are considered to be the North Star when assessing teacher effectiveness. Based on the relationships between value-added scores and the other components in the evaluation model in this study, the other components seem to be valid and reliable measures of teacher effectiveness as well. Although ongoing analyses are needed to ensure continued alignment between the model's components, at present, this study suggests that it is a valuable tool for teacher development and human capital decisions.

The second question of whether school and teacher characteristics are predictive of educator evaluation ratings revealed more interesting findings with additional implications for practice. Because the results showed that certain demographics, such as teacher race and school poverty concentration were in fact predictive of educator ratings, additional analyses should be considered to further investigate the unintended and unmitigated effects of these variables in influencing a teacher's performance. District leaders should consider how evaluation scores are interpreted for certain races of

teachers, particularly when these teachers are serving in more challenging school environments (across-school variance) and serving at-risk populations of students (within-school variance) who oftentimes present additional challenges in classroom settings. While most discussions have raised questions about across-school variance and the impact on teacher performance, fewer studies have raised questions about the impact of within-school variance on teacher performance. Assessing the impact of both factors is particularly important in districts with schools that are serving two different "tracks" of students. If teacher evaluation systems will continue to be used for teacher development and human capital decisions as originally intended, these factors should be examined in depth.

One cautionary tale that must be considered when discussing implications for practice is the context in which he original data were collected. This was not an experimental study in which data were gathered in a controlled setting. These data include real teachers teaching real students in a real urban context. It is important to recognize this exact context when drawing conclusions from the findings, and be explicit about what the findings are and are not saying.

There are a few conclusions to avoid. It should not be assumed that explanations on either polar end are plausible rationale. In other words, it cannot be assumed that all teachers of color are poor or worse teachers, and additionally it should not be assumed that all teacher evaluation models contain implicit bias. Readers should avoid dismissing the TEM as a biased instrument. Instead further investigation should look more closely at race along with more demographic information to find out what other variables might be

at play. Similarly, rather than summarily dismiss TEM as a valid instrument, further study to ensure there is not bias in the tool nor process might be warranted.

One conclusion that is safe to accept is that context matters. As we strive to identify effective teachers, there might be nuance. In other words, when narrowed down to their own population, effective teachers are not widgets either. We know that mutual consent matters, (TNTP, 2009). There might be reasons why effective teachers are more effective in a specific or nuanced context. This has implications for how we think about teacher development and perhaps even more urgently, there might be implications for how we think about student and teacher assignments. The field has mostly treated student and teacher assignments as a random process that does not impact how we think about teacher performance or development. This study might warrant practitioners to ask the question of whether the matching process is indeed random and should continue to be treated as such.

Also, the findings may lead us to think about a more tailored approach for Colleges of Education and alternative certification programs to train teachers for uniquely for urban education, or to work with specific populations or within specific contexts

Conclusion

Nearly a decade ago, the federal government's focus on teacher accountability for student academic outcomes increased. As part of this shift, districts and states began overhauling their teacher evaluation systems to more accurately assess the performance of teachers (Donaldson, 2012). This was a drastic shift from the traditional approach of inconsistently using a sole metric, results of classroom observations, as the measure of

teacher effectiveness. To improve outcomes and earn both federal and private funding, districts across the nation transitioned to educator evaluation models that were publicized for their reliability and predictive ability. Many of these multiple measure models included components in addition to observations, including value-added scores and surveys that measured student perceptions of teacher performance. Although independent of each other, when combined, the components of these evaluation models painted comprehensive pictures of teacher effectiveness. Essentially, these models formed the basis of evaluative and development processes in districts and became the foundation for teacher hiring, retention, compensation and dismissal decisions.

The multiple-measure model employed in the district being considered formed the basis of the analysis in this study. The data revealed that relationships existed between the three primary components of the teacher evaluation system: classroom observations, student perceptions, and value-added, or growth scores. This study also revealed that there were relationships between the components of the evaluation model and characteristics of teachers and schools. In fact, school culture and school poverty concentration were related to the performance ratings that teachers received. Some of the relationships found during the analysis were alarming since evaluation ratings are used for retention and dismissal decisions. One of the most alarming relationships was between observation ratings and school poverty. Teachers serving in the highest poverty schools were rated as less effective than their peers who worked in less challenging environments. If these teachers are actually less effective than their peers in less disadvantaged schools, it is imperative that district leaders focus on strategies to attract

high-quality teachers to the schools where they are most needed. If these teachers are not less effective, however, adjustments should be made to the evaluation system to ensure serving where students' needs are greater does not disadvantage teachers who are delivering high-quality instruction. The results of this study could be used to spearhead conversations with district leaders that focus on the efficacy of the evaluation process for teachers who work in all types of schools and those who serve students from varying backgrounds.

References

Achieve. (2015). Proficient vs. Prepared: Disparities between State Tests and the 2013 National Assessment of Educational Progress (NAEP). Retrieved from http://www.achieve.org/files/NAEPBriefFINAL051415.pdf

Achieve. (2016). Proficient vs. Prepared 2016: State Test Results are Getting Closer to Student Achievement on NAEP. Retrieved from http://www.achieve.org/publications/proficient-vs-prepared-2016

Adams, C. (2015). 2015 SAT, ACT Scores Suggest Many Students Aren't College-Ready. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2015/09/09/2015-sat-act-scores-suggest-many-students.html

Anderson, L. W. (1991). Increasing Teacher Effectiveness. *International Institute for Educational Planning*. Paris: UNESCO.

ASCD (2017). School Culture and Climate. Retrieved from http://www.ascd.org/research-a-topic/school-culture-and-climate-resources.aspx

Belsha, K. (2016). Lowest-scoring teachers concentrated in poorest schools. *The Chicago Reporter*. Retrieved from http://chicagoreporter.com/report-lowest-scoring-teachers-concentrated-in-poorest-schools/

Biddle, B., & Payne, K. (1999). Poor School Funding, Child Poverty, and Mathematics Achievement. *Educational Researcher, 28*, 4-13.

Boser, U., Baffour, P., & Vela, S. (2016). A Look at the Education Crisis. Washington, DC: *Center for American Progress.* Retrieved from https://www.americanprogress.org/issues/education/reports/2016/01/26/129547/a-look-at-the-education-crisis/

Burris, C., & Heubert, J. (2006). Accelerating Mathematics Achievement Using Heterogeneous Grouping. *American Educational Research Journal*, *43*, 105-136.

Camera, L. (2015). Student Scores in Reading and Math Drop. *U.S. News and World Report.* Retrieved from https://www.usnews.com/news/articles/2015/10/28/student-scores-in-reading-and-math-drop

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools. Washington, DC: *Institute of Education Sciences*. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2014024.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2011). The Long-Term Impact of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *National Bureau of Economic Research*. Retrieved from http://www.equality-of-opportunity.org/assets/documents/teachers_wp.pdf

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: *Lawrence Earlbaum Associates.*

Cohen, L., Manion, L., & Morrison, K. (2007). Research Methods in Education. New York: Routledge. Retrieved from https://research-srttu.wikispaces.com/file/view/Research+Methods+in+Education_ertu.pdf

Coleman, J., et al. (1966). Equality of Educational Opportunity. *National Center for Education Statistics.* Retrieved from http://files.eric.ed.gov/fulltext/ED012275.pdf

Connally, K,. & Tooley, M. (2016). Beyond Ratings: Re-envisioning State Teacher Evaluation Systems as Tools for Professional Growth. *New America*. Retrieved from https://static.newamerica.org/attachments/12744-beyond-ratings-3/NA_BeyondRatingsPaper.deba47a82ff04af2833cebdbeed0c3ab.pdf

Dewitt, P. (2013). Classroom Observations are Not About Compliance: What Can We Learn From Failure. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/finding_common_ground/2013/08/what_can_we_learn_from_failure.html?cmp=SOC-SHR-TW

Donaldson, M. (2012). Teachers' Perspectives on Evaluation Perform. *Center for American Progress.* Retrieved from https://cdn.americanprogress.org/wpcontent/uploads/2012/12/TeacherPerspectives.pdf

Dunkin, M. (1997). Assessing Teachers' Effectiveness. *Issues in Educational Research*, 7(1), 37-51. Retrieved from http://www.iier.waier.org.au/iier7/dunkin.html

Ferguson, R. (2014). Tripod Student and Teacher Surveys. *Tripod Education Partners*. Retrieved from https://meeting.nasbonline.org/public/Meeting/Attachments/Display/Attachment.aspx?AttachmentID=122347

Haberman, M. (1995). Star teachers of children of poverty. Indianapolis, IN. In: Kappa Delta Pi.

Hanover Research. (2013). Student Perception Surveys and Teacher Assessments. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiPm8XpqMjSAhWCj1QKHQLSDZEQFgghMAA&url=https%3A%2F%2Fdese.mo.gov%2Fsites%2Fdefault%2Ffiles%2FHanover-Research-Student-Surveys.pdf&usg=AFQjCNFXv59rP1QhT8HRB0ku2WHMOaVu2A

Hanushek, E. (2016). What Matters for Student Achievement? *Education Next*, *16*(2). Retrieved from http://educationnext.org/what-matters-for-student-achievement/

Hinckley, S. (2016). National Report Card Shows Grades are Slipping. Is it Credible? *Christian Science Monitor*. Retrieved from https://www.questia.com/newspaper/1P2-39570722/national-report-card-shows-grades-are-slipping

Houtenville, A., & Conway, K. (2008). Parental Effort, School Resources, and Student Achievement. *The Journal of Human Resources*, *43*, 437-452.

Jiang, J., & Sporte, S. (2016). Teacher Evaluation in Chicago Differences in Observation and Value-Added Scores by Teacher, Student, and School Characteristics. UChicago Consortium on School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/publications/Teacher%20Evaluation%20in%20Chicago-Jan2016-Consortium.pdf

Johnson, C. C. (2007). Effective Science Teaching, Professional Development, and No Child Left Behind: Barriers, Dilemmas, and Reality. *Journal of Science Teacher Education*, *18*(2), 133–136.

Kalogrides, D., Loeb, S., & Beteille, T. (2012). Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education, 86*(2), 103-123. Retrieved from http://asanet.org

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. *Bill and Melinda Gates Foundation*. Retrieved from http://k12education.gatesfoundation.org/%20wp-content/uploads/2015/12/MET_Validating_Using_Random_Assignment_Research_Paper.pdfhttp:/k12education.gatesfoundation.org/%20wp-content/uploads/2015/12/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kena, G., Musu-Gillette, K., Robinson, J., Wang, X., Rathbun, A., Zhang, J., …Ballard, D. (2015). The Condition of Education in 2015. *National Center for Education Statistics*. Retrieved from https://nces.ed.gov/pubs2015/2015144.pdf

Kim, K., & Roth, G. (2011). Novice Teachers and Their Acquisition of Work-Related Information. *Current Issues in Education*, *14*(1). Retrieved from http://cie.asu.edu/

Klein, A. (2015). ESEA Reauthorization: How will ESSA's Regulatory Process Work. *Education Week.* Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2015/12/esea_reauthorization_will_the_.html

Lodico, M., Spaulding, D., & Voegtle, K. (2006). Methods in educational research: From theory to practice. San Francisco: Jossey-Bass. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjFm8vPq8jSAhUKjVQKHYoiDUsQFggcMAA&url=https%3A%2F%2Fwww.researchgate.net%2Ffile.PostFileLoader.html%3Fid%3D56ab012964e9b20c218b4568%26assetKey%3DAS%253A323112883687426%25401454047522725&usg=AFQjCNH6MprVI8LYWDLg7fyj9MXZEvNqoQ

Lubienski, C., & Lubienski, S. (2005). A New Look at Public and Private Schools: Student Background and Mathematics Achievement. *Phi Delta Kappan*, *86*, 696-699.

Magouirk, P. (2014). Understanding TVAAS, Why Student Growth Measures Matter in Tennessee. *SCORE.* Retrieved from http://tnscore.org/understanding-tvaas-why-student-growth-measures-matter-in-tennessee/

Mathis, W. (2006). The Accuracy and Effectiveness of Adequate Yearly Progress, NCLBs School Evaluation System. Retrieved from http://greatlakescenter.org/docs/Policy_Briefs/GLC_AYP_Mathis_FINAL.pdf

McCaffrey, D. (2012). Do Value-Added Methods Level the Playing Field for Teachers. *Carnegie Foundation for the Advanceement of Teaching*. Retrieved from http://files.eric.ed.gov/fulltext/ED537436.pdf

Mead, M. (2007). Easy way out: "Restructured" usually means little has changed. *Education Next*, *7*(1), 52-56.

Mead, S. (2012). Data: Low-Performing Students Disproportionately Assigned to Novice Teachers. *Education Week.*

MET Project. (2010). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. *Bill and Melinda Gates Foundation.*

MET Project. (2012). Asking Students About Teaching: Student Perception Surveys and their Implementation. *Bill and Melinda Gates Foundation.*

MET Project. (2013). Ensuring Fair and Reliable Measures of Effective Teaching. *Bill and Melinda Gates Foundation.*

Musu-Gillette, L. (2015). Education Attainment Differences by Student's Socioeconomic Status. *National Center for Education Statistics.*

National Council of Teachers of English. (2008). English Language Learners. Urbana, Illinois.

Papay, J. (2013). What's Driving Teachers Away from High-Poverty Schools. *American Institutes for Research.* Retrieved from http://www.gtlcenter.org/blog/what%E2%80%99s-driving-teachers-away-high-poverty-schools

Partee, G. (2012). Using Multiple Evaluation Measures to Improve Teacher Effectiveness. Washington, DC: *Center for American Progress*. Retrieved from https://www.americanprogress.org/issues/education/reports/2012/12/18/48368/using-multiple-evaluation-measures-to-improve-teacher-effectiveness/

Ramona, S. (2011). Advantages and Disadvantages of Quantitative and Qualitative Information Risk Approaches. *Chinese Business Review*, *10* (12), 1106-1110. Retrieved from http://www.davidpublishing.com/davidpublishing/Upfile/1/6/2012/2012010671957745.pdf

Ravitch, D. (2015). Diane Ravitch's Blog: Senate Committee Reaches Agreement on New ESEA. Retrieved from http://nepc.colorado.edu/blog/senate-committee

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. Retrieved from https://econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf

Sawchuk, S. (2016) ESSA Loosens Reins on Teacher Evaluations, Qualifications. *Education Week.*

Shelby County Schools. (2013). Teacher Effectiveness. Retrieved from http://www.scsk12.org/Policy_Manual/pm/4000/4045_Teacher_Effectiveness.pdf

Simon, M. (2011). Dissertation and Scholarly Research: Recipes for Success. Seattle, WA: Dissertation Success LLC.

Skinner, R. (2009). The No Child Left Behind Act: An Overview of Reauthorization Issues for the 111th Congress. *Congressional Research Service*. Retrieved from http://www.leahy.senate.gov/imo/media/doc/The%20No%20Child%20Left%20Behind%20Act%20-%20An%20Overview%20of%20Reauthorization%20Issues%20for%20the%2011th%20Congress.pdf

Sparks, S. (2016). Study: Most School Districts have Achievement Gaps. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2016/05/11/study-most-school-districts-have-achievement-gaps.html

Special Education Guide. (2016). What is Special Education? Retrieved from http://www.specialeducationguide.com/pre-k-12/what-is-special-education/

Taylor, J. (2005). Poverty and Student Achievement. *Multicultural Education, 12* (4), 53-55.

Tennessee Department of Education. (2012). *Teacher Evaluation in Tennessee: A Report on Year 2 Implementation.* Retrieved from http://team-tn.org/wp-content/uploads/2013/08/yr_2_tchr_eval_rpt.pdf

The Education Trust. (2016). *The Every Student Succeeds Act: What's In It? What Does it Mean for Equity.* Retrieved from https://edtrust.org/resource/the-every-student-succeeds-act-whats-in-it-what-does-it-mean-for-equity/

The Nation's Report Card (2015). Retrieved from https://www.nationsreportcard.gov/reading_math_2015/#?grade=4

The New Teacher Project. (2010). *Teacher Evaluation 2.0.* Brooklyn, NY.

The Social Welfare History Project (2014). Elementary and Secondary Education Act of 1965. Retrieved from http://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/

TNTP. (2012). *Greenhouse Schools*. Brooklyn, NY.

TNTP. (2013). *Fixing Classroom Observations*. Brooklyn, NY.

Turner, C. (2015). No Child Left Behind: What Worked, What Didn't. *nprED.*

University of Chicago Consortium of School Research. (2016). Teachers with lowest evaluation scores are overrepresented in high poverty schools. Retrieved from https://consortium.uchicago.edu/sites/default/files/pressreleases/REACH%20III%20Press%20Release.pdf

U.S. Department of Education. (2009). American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education. Retrieved from https://www2.ed.gov/policy/gen/leg/recovery/implementation.html

Walker, T. (2015) With Passage of Every Student Succeeds Act, Life After NCLB Begins. *neaTODAY*. Retrieved from http://neatoday.org/2015/12/09/every-student-succeeds-act/

Weisberg, D., Sexton, S., Mulhern, J. & Keeling, D. (2009). The Widget Effect. The New Teacher Project. Brooklyn, NY.

Whitehurst, G., Chingos, M., & Lindquist, K. (2014). Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts. *Brown Center on Education Policy at Brookings.* Retrieved from https://www.brookings.edu/research/evaluating-teachers-with-classroom-observations-lessons-learned-in-four-districts/

Wirt, J., Choy, S., Provasnik, S., Rooney, P., Sen, A., & Tobin, R. (2003). The Condition of Education 2003. *National Center for Education Statistics.* Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjFlPzcsMjSAhUI4SYKHReSADcQFggeMAA&url=https%3A%2F%2Fnces.ed.gov%2Fpubs2004%2F2004077.pdf&usg=AFQjCNFeK6BzKySsih0vzsZnWQN-Glm3Jg

Zernike, K. (2016, August 27). Test Scores Show a Decline in Math Among High School Seniors. *New York Times*. Retrieved from https://www.nytimes.com/2016/04/27/us/math-test-scores-decline-high-school-seniors.html?_r=0